

# Robust Text Independent Closed Set Speaker Identification Systems and Their Evaluation



Musab Tahseen Salahaldeen Al-Kaltakchi

Newcastle University

Newcastle Upon Tyne, UK

A thesis submitted for the degree of

*Doctor of Philosophy*

January 2018

## CERTIFICATE OF ORIGINALITY

I, Musab Tahseen Salahaldeen Al-Kaltakchi, confirm that the work in this thesis is my own. I have read and understand the penalties associated with plagiarism.

Signature:

Date:

## SUPERVISOR'S CERTIFICATE

This is to certify that the entitled thesis “Robust Text Independent Closed Set Speaker Identification Systems and Their Evaluation” has been prepared under my supervision within the School of Electrical and Electronic Engineering / Newcastle University for the degree of PhD in Electrical Engineering / Digital Signal Processing.

Signature:

Supervisor: Prof. Dr. Jonathon A. Chambers

Date:

Signature:

Student: Musab Tahseen Salahaldeen Al-Kaltakchi

Date:



Surah ar-Room (The Romans)

وَمِنْ آيَاتِهِ خَلْقُ السَّمَاوَاتِ وَالْأَرْضِ وَاخْتِلَافُ  
الْسِّنَتِكُمْ وَأَلْوَانِكُمْ إِنَّ فِي ذَلِكَ لَآيَاتٍ لِّلْعَالَمِينَ

[30:22] And one of His signs is the creation of the heavens and the earth and the diversity of your tongues and colors; most surely there are signs in this for the learned.

صَدَقَ اللَّهُ الْعَظِيمُ



# Abstract

This thesis focuses upon text independent closed set speaker identification. The contributions relate to evaluation studies in the presence of various types of noise and handset effects. Extensive evaluations are performed on four databases.

The first contribution is in the context of the use of the Gaussian Mixture Model-Universal Background Model (GMM-UBM) with original speech recordings from only the TIMIT database. Four main simulations for Speaker Identification Accuracy (SIA) are presented including different fusion strategies: Late fusion (score based), early fusion (feature based) and early-late fusion (combination of feature and score based), late fusion using concatenated static and dynamic features (features with temporal derivatives such as first order derivative delta and second order derivative delta-delta features, namely acceleration features), and finally fusion of statistically independent normalized scores.

The second contribution is again based on the GMM-UBM approach. Comprehensive evaluations of the effect of Additive White Gaussian Noise (AWGN), and Non-Stationary Noise (NSN) (with and without a G.712 type handset) upon identification performance are undertaken. In particular, three NSN types with varying Signal to Noise Ratios (SNRs) were tested corresponding to: street traffic, a bus interior and a crowded talking environment. The performance evaluation also considered the effect of late fusion techniques based on score fusion, namely mean, maximum, and linear weighted sum fusion. The databases employed were: TIMIT, SITW, and NIST 2008; and 120 speakers were selected from each database to yield 3,600 speech utterances.

The third contribution is based on the use of the I-vector, four combinations of I-vectors with 100 and 200 dimensions were employed. Then, various fusion techniques using maximum, mean, weighted sum and cumulative fusion with the same I-vector dimension were used to improve the SIA. Similarly, both interleaving and concatenated I-vector fusion were exploited to produce 200 and 400 I-vector dimensions. The system was evaluated with four different databases using 120 speakers from each database. TIMIT, SITW and NIST 2008 databases were evaluated for various types of NSN namely, street-traffic NSN, bus-interior NSN and crowd talking NSN; and the G.712 type handset at 16 kHz was also applied.

As recommendations from the study in terms of the GMM-UBM approach, mean fusion is found to yield overall best performance in terms of the SIA with noisy speech, whereas linear weighted sum fusion is overall best for original database recordings. However, in the I-vector approach the best SIA was obtained from the weighted sum and the concatenated fusion.

# Statement of Originality

The contributions of this thesis have been supported by different journal and conference papers, which have been generated during the journey of my study. They can be stated as follows:

Part A: Published papers (conferences):

[1] M.T.S. Al-Kaltakchi, W.L. Woo, S.S. Dlay, and J.A. Chambers, “Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification,” in *4th IEEE International Conference on Biometrics and Forensics (IWBF)*, Limassol, Cyprus, 2016, pp. 1-6.

DOI: 10.1109/IWBF.2016.7449685.

[2] M.T.S. Al-Kaltakchi, W.L. Woo, S.S. Dlay, and J.A. Chambers, “Study of statistical robust closed set speaker identification with feature and score-based fusion,” in *2016 IEEE Statistical Signal Processing Workshop (SSP), Palma de Mallorca, Spain, 2016*, pp. 1-5.

DOI: 10.1109/SSP.2016.7551807.

[3] M.T.S. Al-Kaltakchi, W.L. Woo, S.S. Dlay, and J.A. Chambers, “Speaker identification evaluation based on the speech biometric and I-vector model using the TIMIT and NTIMIT databases,” in *5th IEEE International Workshop on Biometrics and Forensics (IWBF)*, Coventry, UK, 2017, pp. 1-6.

DOI: 10.1109/IWBF.2017.7935102.

The contribution of paper [1] in terms of the GMM-UBM approach and combining PNCC and MFCC features provides a robust performance for original speech recordings and noisy environments. It achieves enhancement in SIA by removing and reducing sensitivity due to the channel between the speaker and microphone together with handsets by using normalization methods, feature warping and CMVN. Moreover, instead of modelling individual speakers with limited data

only by a GMM as in a previous study, a GMM-UBM based modelling strategy is used as in [2] utilizing all speakers data to increase the number of mixtures and thereby enhance the identification rate. Furthermore, weighted sum, maximum and mean fusion are studied with the combination of the features scores as methods to improve SIA. This contribution is explained in Chapter 4 in this thesis.

The major contribution for paper [2] is also in terms of the GMM-UBM approach and is to perform a thorough evaluation of the scheme by conducting more sophisticated fusion schemes in the presence of the handset and AWGN. Furthermore, exploitation of fusion techniques depending on feature dimension is considered such as early feature fusion (32 feature dimension), late score fusion (16 feature dimension) and finally combination of feature and score based early and late fusion (32 feature dimension). A 4th order G.712 type IIR filter is employed to represent handset degradation in the channel. This contribution can be found in Chapter 4 for different fusion schemes tested under original speech recordings. However, the evaluations in the presence of handset and AWGN effects are given in Chapter 5.

The contribution of paper [3] is in terms of the I-vector approach for speaker identification purpose. The system is modelled by I-vectors with a fixed dimension of 100 and the performance accuracy is improved using maximum, mean and weighted sum fusion. Then the extreme learning machine is utilized for classification purposes. Furthermore, a fair comparison is performed using both TIMIT and NTIMIT databases. The paper presents new combinations of MFCC and PNCC features modelled by I-vectors, then the fusion based I-vectors are classified by using the ELM to improve the Speaker Identification Accuracy (SIA). This contribution is presented in Chapter 6.

[4] M.T.S. Al-Kaltakchi, W.L. Woo, S.S. Dlay, and J.A. Chambers, “Comparison of I-vector and GMM-UBM Approaches to Speaker Identification with TIMIT and NIST 2008 Databases in Challenging Environments,” *in 25th European Signal Processing Conference (EUSIPCO)*, 2017, Kos, Greece.

DOI: 10.23919/EUSIPCO.2017.8081264.

The main contribution for paper [4] is a comparison of the I-vector and the Gaussian Mixture Model-Universal Background Model (GMM-UBM), approaches for the speaker identification task. Four feature combinations of I-vectors with seven fusion techniques are considered. In addition, an Extreme Learning Machine (ELM) is exploited to identify speakers, and then SIA is calculated. Both systems are evaluated for 120 speakers from the TIMIT and NIST 2008 databases for original speech recordings. Furthermore, a comprehensive evaluation is made with AWGN conditions and with three types of NSN. The contribution in terms of the GMM-UBM is explained in Chapter 5 in this thesis, while the part related with the I-vector is explained in Chapter 6.

Part B: Accepted paper (conference):

[5] M.T.S. Al-Kaltakchi, W.L. Woo, S.S. Dlay, and J.A. Chambers, “Multi-dimensional I-vector closed set speaker identification based on an extreme learning machine with and without fusion technologies,” *Accepted in Intelligent Systems Conference (IntelliSys)* , 2017, London, UK.

The main contribution for paper [5] is to investigate, regardless of channel variability, a new speaker model representation based on the I-vector which exploits the ELM for relatively low complexity and outperforms the traditional GMM-UBM for modelling speakers in the speaker identification task in both original speech recordings and AWGN environments. Furthermore, a smaller size I-vector with (100, 200) dimensions and multidimensional I-vectors are presented in this

paper using the I-vector with dimensions of 200, 400 and 800 to improve the identification accuracy. This work is explained in Chapter 6.

Part C: Published Journal:

[6] M.T.S. Al-Kaltakchi, W.L. Woo, S.S. Dlay, and J.A. Chambers, “Evaluation of a Speaker Identification System With and Without Fusion Using Three Databases in the Presence of Noise and Handset Effects”, *EURASIP Journal on Advances in Signal Processing*, Processing (2017) 2017:80. DOI: 10.1186/s13634-017-0515-7.

<https://doi.org/10.1186/s13634-017-0515-7>.

This article has three contributions related to GMM-UBM approach: firstly a speaker identification system was established using different feature combinations based on MFCC and PNCC features and comprehensive evaluations of the effect of both AWGN, and NSN (with and without a G.712 type handset) upon identification performance were undertaken. In particular, three NSN types with varying SNRs were tested corresponding to: street traffic, a bus interior and a crowded talking environment. Secondly, three databases were employed: TIMIT, SITW, and NIST 2008; and 120 speakers were selected from each database to yield 3,600 speech utterances for more extensive evaluations. Thirdly, different late fusion methods were used to improve the SIA. Details of this paper can be found in Chapter 5.

Part D: Submitted Journal:

[7] M.T.S. Al-Kaltakchi, W.L. Woo, S.S. Dlay, and J. A. Chambers, “Combined I-vector and Extreme Learning Machine Approach for Robust Speaker Identification and Evaluation with SITW 2016, NIST 2008, TIMIT Databases”, *Submitted to IET Biometrics, 2017.*

This paper has three contributions related to comprehensive evaluations in terms of the I-vector approach. Firstly, a new text independent closed set speaker identification system which exploits four feature combinations of I-vectors with 100 and 200 I-vector dimensions and seven fusion techniques is considered. Secondly, the system is evaluated on three different databases with 120 speakers from each database. Thirdly, the system was evaluated under original speech recordings, AWGN and NSN namely, street-traffic, bus-interior and crowd talking. This work is described in Chapter 6.

Additional papers has been established during my PhD journey.

[8] R. R. O. Al-Nima, M. A. M. Abdullah, M. T. S. Al-Kaltakchi, S. S. Dlay, W. L. Woo, and J. A. Chambers, “Finger Texture Biometric Verification Exploiting a Multi-scale Sobel Angles Local Binary Pattern and Score-based Fusion”, *Elsevier, Digital Signal Processing*, vol. 70, 2017, Pages 178-189.

<https://doi.org/10.1016/j.dsp.2017.08.002>

[9] R. R. O. Al-Nima, M. T. S. Al-Kaltakchi, S. S. Dlay, W. L. Woo, and J. A. Chambers, “Personal Verification Based on Multi-Spectral Finger Texture Lighting Images”, *Submitted to IET Signal Processing, 2017.*

## Acknowledgements

I am deeply and sincerely indebted to my main supervisor Professor Jonathon Chambers for his consistent instruction, constant support and generous advice throughout my PhD period. I have benefited tremendously from his rare insight, his exceptional knowledge and his great enthusiasm to students. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly. I have learned much from his instruction. It is my great privilege and exclusive honour to have been one of his research students. I wish that I will have more opportunities to work with him in the future.

I am also extremely thankful to Dr Wai Lok Woo the second supervisor and Professor Satnam Dlay the third supervisor for their support through my PhD period.

I also wish to gratefully thank the Ministry of Higher Education and Scientific Research (MoHESR), and the Iraqi Cultural Attaché, Al-Mustansiriya University, Al-Mustansiriya University College of Engineering in Iraq for supporting my PhD scholarship.

I also want to record my appreciation to all friends, relatives and others who in one way or another shared their support either morally, physically or financially.

I also gratefully thank my parents for their encouragements to complete my PhD study. I hope for a very long and happy life for my Mother. I pray to my father although he died during my first year of PhD study, I hope his spirit will be forgiven and he will rest in peace. I did not visit him in his last place but I intend to visit him and carry to him a doctorate that he so wished to see me achieve before his death.

I am also extremely thankful to all my brothers especially my big



brother Basil Tahseen Salahaldeen AlKaltakchi for all their support through out my life and my PhD study.

I also gratefully thank my lovely family especially my lovely son YOUSIF and my wife for all their support through all of my life and through my PhD.

Musab Al-Kaltakchi

September 2017

# Contents

<b>List of Acronyms</b>	<b>xxviii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background to Speaker Identification . . . . .	1
1.2 Aims . . . . .	5
1.3 Motivation . . . . .	6
1.4 Objectives . . . . .	6
1.5 Difficulties with a Speaker Identification System . . . . .	6
1.6 Speaker Recognition Applications . . . . .	8
1.7 Thesis structure . . . . .	9
<b>2 Background and Related Literature Review</b>	<b>11</b>
2.1 The auditory perception system . . . . .	11
2.2 Background to Speaker Identification Systems . . . . .	14
2.2.1 Feature Extraction . . . . .	15
2.2.2 Mel Frequency Cepstral Coefficients (MFCCs) . . . . .	15
2.2.3 Power Normalized Cepstral Coefficients (PNCCs) . . . . .	18
2.2.4 Comparisons Between MFCC and PNCC Features . . . . .	20
2.2.5 Feature Normalization Methods . . . . .	21
2.3 Literature Review for Speaker Identification . . . . .	22
2.3.1 Feature Extraction . . . . .	22
2.3.2 Literature Review in Terms of I-vector Extraction . . . . .	27
2.3.3 Modelling . . . . .	28
2.3.4 Noise Robustness and Challenging Environments . . . . .	33
2.3.5 Classification . . . . .	38
2.3.6 Fusion Technologies . . . . .	39
2.4 Summary . . . . .	41

<b>3</b>	<b>Databases and Performance Measurement</b>	<b>43</b>
3.1	Background . . . . .	43
3.2	Databases . . . . .	43
3.2.1	Type 1: Databases Used in This Thesis . . . . .	48
3.2.1.1	TIMIT Acoustic-Phonetic Continuous Speech Corpus-1993 . . . . .	48
3.2.1.2	NTIMIT-Network TIMIT . . . . .	50
3.2.1.3	The Speakers In The Wild Speaker Recognition Challenge 2016 . . . . .	52
3.2.1.4	2008 NIST Speaker Recognition Evaluation Training Set Part 2-2011 . . . . .	52
3.2.1.5	Non Stationary Noise Database . . . . .	54
3.2.2	Type 2: Databases Not Used in This Thesis . . . . .	54
3.2.2.1	MOBIO Database . . . . .	54
3.2.2.2	The GRID audiovisual sentence corpus . . . . .	55
3.2.2.3	VoxForge Database . . . . .	55
3.2.2.4	YOHO Database . . . . .	56
3.2.2.5	NIST I-vector Machine Learning Challenge Databases	56
3.2.2.6	MATLAB Audio Databases Toolbox . . . . .	56
3.3	Performance Measurement . . . . .	57
3.3.1	Part A: DET Curve, EER and min DCF . . . . .	57
3.3.2	Part B: Speaker Identification Accuracy . . . . .	59
3.4	Summary . . . . .	59
<b>4</b>	<b>Speaker Identification Using GMM-UBM Approach With Fusion and Evaluated on Original Speech Recordings</b>	<b>60</b>
4.1	Background . . . . .	61
4.2	Biometric Speaker Identification Framework . . . . .	63
4.2.1	Feature Extraction and Feature Normalization . . . . .	63
4.2.2	Acoustic Modelling and Matching . . . . .	64
4.2.2.1	Universal Background Model and GMM-UBM . . . . .	64
4.2.2.2	Adaptation of Speaker Models . . . . .	65
4.2.2.3	Maximum Log-likelihood Scores . . . . .	66
4.3	Speaker Identification Systems With Fusion Strategies . . . . .	69

4.3.1	System 1: Speaker Identification System With Late Fusion . .	69
4.3.2	System 2: Speaker Identification System With Early Fusion and Early-Late Fusion . . . . .	72
4.3.3	System 3: Speaker Identification System With Late Fusion for Concatenated Static and Dynamic Features . . . . .	74
4.3.4	System 4: Speaker Identification System With Late Fusion For Normalized Independent Scores For Systems 1, 2 and 3 . .	75
4.4	Simulations Setup . . . . .	76
4.5	Related Work . . . . .	77
4.6	Simulation Results . . . . .	78
4.6.1	Simulation Results For System 1 . . . . .	78
4.6.2	Simulation Results For System 2 . . . . .	80
4.6.3	Simulation Results For System 3 . . . . .	82
4.6.4	Simulation Results For System 4 . . . . .	82
4.7	Discussions . . . . .	84
4.8	Summary . . . . .	88
<b>5</b>	<b>Speaker Identification Using GMM-UBM Approach With Fusion For Challenging Environments With Three Databases</b>	<b>91</b>
5.1	Background . . . . .	92
5.2	An Overview of a Robust Biometric Speaker Identification System .	94
5.2.1	Feature Extraction and Compensation . . . . .	95
5.2.2	Speaker Modelling and Matching . . . . .	95
5.2.2.1	Gaussian Mixture Model (GMM) . . . . .	95
5.2.2.2	Gaussian Mixture Model-Universal Background Model (GMM-UBM) . . . . .	96
5.2.2.3	Maximum Log-Likelihood Scores . . . . .	96
5.2.3	Fusion Strategies . . . . .	97
5.3	Adding Noise and Applying The G.712 Type Handset . . . . .	97
5.3.1	Adding Stationary AWGN and Non-Stationary Noise . . . . .	97
5.3.2	G.712 Type Handset . . . . .	97
5.4	Databases and Simulation Setups . . . . .	99
5.4.1	Databases . . . . .	99

5.4.1.1	TIMIT Acoustic-Phonetic Continuous Speech Corpus-1993 . . . . .	99
5.4.1.2	The Speakers In The Wild Speaker Recognition Challenge 2016 . . . . .	99
5.4.1.3	2008 NIST Speaker Recognition Evaluation Training Set Part 2-2011 . . . . .	99
5.4.2	Simulation Setups . . . . .	101
5.5	Simulation Results and Discussion . . . . .	101
5.5.1	Simulations and Experiments for Part A . . . . .	102
5.5.1.1	Evaluation of Speech Data from TIMIT, SITW and NIST 2008 Without Handset and Noise (Part A) . .	102
5.5.1.2	Evaluation of Noisy Speech Data from TIMIT, SITW and NIST 2008 With Handset and Noise (Part A) . . . . .	104
5.5.2	Simulations and Experiments for Part B . . . . .	112
5.5.2.1	Quantitative Perspective for Noise and Handset Effects in PartB . . . . .	112
5.6	Related Works Based on the Proposed Speaker Identification System	113
5.7	Summary . . . . .	114
<b>6</b>	<b>Fusion-based Speaker Identification Using Multi-Dimensional I-vectors in Challenging Environments for Four Databases</b>	<b>117</b>
6.1	Background . . . . .	118
6.2	Related Work . . . . .	121
6.3	Fusion-based I-vector scheme . . . . .	123
6.3.1	Compact features extraction and normalization . . . . .	124
6.3.2	I-vectors extraction framework . . . . .	124
6.3.3	Fusion Methods Based on I-vectors . . . . .	126
6.3.4	ELM classification and calculating the identification accuracy	129
6.4	Simulation Setups . . . . .	131
6.4.1	Databases and Environments . . . . .	131
6.5	Experimental Results and Discussions . . . . .	133
6.5.1	The Relationship Between Multi-Dimensional I-vectors and SIA in TIMIT Database Evaluations . . . . .	139

6.5.2	The Effects of Mixture Sizes, SNR Levels and Feature Combinations of I-vectors Without/With Fusion for SIA in TIMIT Database Evaluations . . . . .	141
6.5.3	Comparisons of I-vector and GMM-UBM Approaches in Terms of The Speaker Identification Accuracy . . . . .	144
6.6	Recent Works Related to I-vector and GMM-UBM Techniques Speaker Identification . . . . .	149
6.7	Summary . . . . .	152
6.8	<b>Appendix 6.1</b> . . . . .	155
<b>7</b>	<b>Conclusions and Future Work</b>	<b>176</b>
7.1	Contributions Overview . . . . .	176
7.2	Conclusions . . . . .	178
7.3	Suggestions for future work . . . . .	186
	<b>References</b>	<b>188</b>

# List of Figures

1.1	Block Diagram of a Speaker Identification System [3] . . . . .	2
1.2	The Speaker Recognition Tasks: (A) Speaker, Language and Speech Recognition; (B) Task 1: Closed Set Speaker Identification; (C) Task 2: Open Set Speaker Identification; (D) Task 3: Verification, Detection and Authentication . . . . .	3
1.3	Block Diagram of a Speaker Verification System [3] . . . . .	4
2.1	The Main Speaker Identification Scheme in This Thesis . . . . .	12
2.2	Human Communication System [8] . . . . .	12
2.3	A Cross Section of the Human Ear [9] . . . . .	13
2.4	Implant Cochlea [10] . . . . .	14
2.5	Mel Frequency Cepstral Coefficients Features (MFCCs) [3] [13] . . . .	17
2.6	Power Normalized Cepstral Coefficients (PNCCs) Features [15] . . . .	19
2.7	Comparison Between MFCC and PNCC Features Structure [6] . . . .	20
2.8	Percentage Proportion of Each Stage of the Speaker Identification System in 54 Studied in the Literature Review Based on Six Stages of the System: Feature Extraction, I-vector Extraction, Modelling, Noise Robustness, and Challenging Environments, Classifiers and Fusion Techniques. . . . .	42
4.1	Trials Production for 120 Speakers From TIMIT Database with 120 speaker model and Four Testing Utterances Per Speaker (Total 480 Testing) to Yield 57,600 Trials of Model-Test Sets . . . . .	68
4.2	Flowchart for Speaker Identification System Multi-Bases (16D) With/Without (W/WO) Late Fusion . . . . .	71
4.3	Flowchart for Speaker Identification System Multi-Bases (32D) Early Fusion With/Without Late Fusion . . . . .	73

## LIST OF FIGURES

---

4.4	Concatenated MFCC/PNCC Static Features with the Dynamic Features with 39 Dimension [3] . . . . .	75
4.5	System 4 Independent Scores for Different Feature Dimensions . . . .	77
4.6	Box Plot of SIA for Original Speech Recordings with Multi-Bases and Late Fusion Proposed Algorithms for 16D and Relate to Simulation 1, with represents $\omega_\beta$ Weights . . . . .	84
4.7	Ribbon Plot for Original Speech Recordings Based on Late Fusion Approaches for 16D and Relate to Simulation 1 . . . . .	85
4.8	Comparison SIA in Original Speech Recordings Based on Early Fusion (Multi-Bases Lines) with Early Late Fusion Relate to simulation 2 . .	86
4.9	Comparison of the Performance for Original Speech Recordings Between Late Fusion (simulation 1) with the Early-Late Fusion (simulation 2) . . . . .	87
4.10	SIA Against GMCs for ALL Original Speech Recordings Simulations	88
5.1	Robust Biometric Speaker Identification and Evaluation Framework. .	94
5.2	Frequency Response for G.712 Type Handset . . . . .	98
5.3	Impulse Response for G.712 Type Handset . . . . .	98
5.4	Evaluations in Terms of SIA for the TIMIT, SITW and NIST 2008 Databases for Widespread Gaussian Mixture Components {8, 16, 32, 64, 128, 256, 512} Without Handset and Noise Using the GMM-UBM Algorithm . . . . .	104
5.5	Performance Measurement for Noisy Speech for the TIMIT, SITW and NIST 2008 Database at Mixture Size 256 Under G.712 Type Handset at 16 kHz With Background Noise (a) AWGN , (b) Street Traffic NSN, (c) Bus-Interior and (d) Crowd Talking NSN for Wide Range of SNR Levels (0-30) dB and Using GMM-UBM Algorithm. . .	110
6.1	Text-Independent Speaker Identification Scheme . . . . .	123
6.2	I-vector Extraction Block Diagram . . . . .	125
6.3	I-vector Fusion Scheme Methods: (a) Maximum, (b) Mean, (c) Cumulative and (d) Weighted Sum (with d-Dimension), (e) Concatenated with 2d and 4d-Dimensions, and (f) Interleaving Fusion (with 2d-Dimension) . . . . .	127



6.4	Structure of Single Layer Feedforward Extreme Learning Machine with Input Dimension $d$ , $L$ hidden nodes and $L$ outputs [159]	130
6.5	The Highest SIA Using the I-vector with 100, 200 and 400 Dimensions per UBM Mixture Size for TIMIT, SITW, NIST2008 and NTIMIT Databases; which Represented the Best SIA for the Tables: Table 6.3, Table 6.13, Table 6.18 and Table 6.23, Respectively	134
6.6	The Bar Chart Shows the Highest SIA at each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT, SITW and NIST2008 Databases Under AWGN with G.712 Type Handset at 16 kHz; the Best SIAs are found in: Table 6.9, Table 6.14 and Table 6.19, Respectively	136
6.7	The Bar Chart Shows the Highest SIA at Each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT, SITW and NIST2008 Databases Under Street Traffic NSN with G.712 Type Handset at 16 kHz; the Best SIAs are found in: Table 6.10, Table 6.15 and Table 6.20, Respectively	137
6.8	The Bar Chart Show the Highest SIA for Each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT, SITW and NIST2008 Databases Under Bus Interior NSN with G.712 Type Handset at 16 kHz; the Best SIAs are found in: Table 6.11, Table 6.16 and Table 6.21, Respectively	138
6.9	The Bar Chart Show the Highest SIA for Each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT, SITW and NIST2008 Databases Under Crowd Talking NSN with G.712 Type Handset at 16 kHz; the Best SIAs are found in: Table 6.12, Table 6.17 and Table 6.22, Respectively	139
6.10	The Bar Chart Shows the Highest SIA for Each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT Database Under AWGN, Street Traffic, Bus Interior and Crowd Talking NSN without Handset; the Best SIAs are found in: Table 6.5 to Table 6.8, Respectively	140
6.11	The Relationship Between SIA and Multi-Dimensional I-vectors for Different UBM Mixture Sizes for the original speech recordings of TIMIT Database	141

6.12	Box Plots for TIMIT Database Evaluation in original speech recordings and AWGN Noisy Speech Based on I-vector With/Without (W/WO) Fusion : Simulation 1 Represented by Part (a) and Part (c); Simulation 2 Represented by Part (b) and Simulation 3 Represented by Part (d) and Part (e): where $\mathbf{f}_1$ and $\mathbf{f}_2$ , $\mathbf{g}_1$ and $\mathbf{g}_2$ are FWMFCC, CMCNMFCC, FWPNCC and CMVNPCC I-vector Features with d-Dimension; Fusion Sets Symbols $\mathbf{F}_1$ , $\mathbf{F}_2$ , $\mathbf{F}_3$ are d-Dimension I-vectors for Weighted Sum, Maximum and Mean Fusion. $\mathbf{F}_4$ is d-Dimension Cumulative Fusion I-vector, $\mathbf{F}_5$ and $\mathbf{F}_6$ are Concatenated and Interleaving Fusion I-vectors with 2d Dimension, $\mathbf{F}_7$ is Concatenated Fusion for the Four Feature Combinations of the I-vectors with 4d-Dimension. . . .	142
6.13	Comparison of Two Speaker Identification Frameworks Using GMM-UBM and I-vector Approaches Evaluated Under Different Environmental Conditions: Part A, Training Phase; Part B, Testing Phase . . . . .	145
6.14	Bar Chart Plot Comparisons Between SIA Against Gaussian Mixture Components for GMM-UBM and I-vector Approaches in Terms of Original Speech Recordings From TIMIT Database . . . . .	146
6.15	Curve Plot Comparison GMM-UBM and I-vector Approaches for AWGN and NSN without Handset at UBM Mixture Size 256 for TIMIT Database . . . . .	147
6.16	Curve Plot Comparison GMM-UBM and I-vector Approaches for AWGN and NSN with G.712 Type Handset at 16 kHz at UBM Mixture Size 256 for TIMIT Database . . . . .	148
6.17	The Comparison Between the SIA for the GMM-UBM and I-vector Approaches for Speech Utterances From SITW Database without Noise and Handset . . . . .	149
6.18	The Comparison Between the SIA for the GMM-UBM and I-vector Approaches for AWGN, Street Traffic NSN, Bus Interior NSN and Crowd Talking NSN with G.712 Type Handset at 16 kHz (at UBM mixture size 256 ) for The SITW Database . . . . .	150

## LIST OF FIGURES

---

6.19 The Comparison Between the SIA for the GMM-UBM and I-vector Approaches for Speech Utterances From NIST 2008 Database without Noise and Handset . . . . .	151
6.20 The Comparison Between the SIA for the GMM-UBM and I-vector Approaches for AWGN, Street Traffic NSN, Bus Interior NSN and Crowd Talking NSN with G.712 Type Handset at 16 kHz (at UBM Mixture Size 256 ) for the NIST 2008 Database . . . . .	151

# List of Tables

3.1	Description of the TIMIT Family of Databases-Part A . . . . .	45
3.2	Description of the TIMIT Family of Databases-Part B . . . . .	46
3.3	Description of the TIMIT Family of Databases -Part C . . . . .	47
3.4	TIMIT Dialect Region Distribution of Speakers . . . . .	49
3.5	Speech Material for TIMIT Corpus . . . . .	49
3.6	The NTIMIT Database Files Lacking Speech Data . . . . .	50
3.7	NTIMIT Database Files Lacking Calibration Data . . . . .	51
4.1	Experimental Parameters for the Work in [1] and in All Proposed Simulations in This Chapter . . . . .	78
4.2	Main Comparison Between the Work in [1] and the Proposed Algorithm . . . . .	79
4.3	SIA Results for Original Speech Recordings as in [1] . . . . .	79
4.4	Simulation 1: Speaker Identification System with Late Fusion . . . . .	80
4.5	Simulation 2: Speaker Identification System with Early Fusion and Early-Late Fusion . . . . .	81
4.6	Simulation 3: Speaker Identification System with Late Fusion for the Concatenated of the Static and Dynamic Features . . . . .	82
4.7	Simulation 4: Speaker Identification System with Late Fusion for Normalized Independent Scores for Systems 1, 2 and 3 . . . . .	83
5.1	Parameters and setup used in all experiments and simulations . . . . .	100
5.2	Simulation 1: 1 A, 1 B and 1 C are the SIA for Different Gaussian Mixture Components (GMC) for the TIMIT, SITW and NIST 2008, Respectively . . . . .	105

## LIST OF TABLES

---

5.3	Simulation 2: 2 A, 2 B and 2 C are the SIA Under AWGN and G.712 Type Handset at 16 kHz for Different Signal to Noise Ratio (SNR) Levels for the TIMIT, SITW and NIST 2008, Respectively, at Mixture Size 256 . . . . .	106
5.4	Simulation 3: 3 A, 3 B and 3 C are the SIA for Street Traffic NSN and G.712 Type Handset at 16 kHz for Different Signal to Noise Ratio (SNR) Levels for the TIMIT, SITW and NIST 2008, Respectively, at Mixture Size 256 . . . . .	107
5.5	Simulation 4: 4 A, 4 B and 4 C are the SIA for Bus Interior NSN and G.712 Type Handset at 16 kHz for Different Signal to Noise Ratio Levels for the TIMIT, SITW and NIST 2008, Respectively, at Mixture Size 256 . . . . .	108
5.6	Simulation 5: 5 A, 5 B and 5 C are The SIA for Crowded Talking NSN and G.712 Type Handset at 16 kHz for Different Signal to Noise Ratio Levels for the TIMIT, SITW and NIST 2008, Respectively, at Mixture Size 256 . . . . .	109
5.7	Percentage Reduction in SIA (PRSIA) for the TIMIT, SITW and NIST 2008, Respectively, Under G.712 Type Handset at 16 kHz at Mixture Size 256 and SNR 30 dB, AWGN, Street Traffic, Bus Interior, Crowded Talking NSN . . . . .	114
5.8	Comparisons with the State of the Art of SIA . . . . .	116
6.1	Parameters and Setup Used in All Experiments and Simulations . . .	132
6.2	Related Work with I-vector and GMM-UBM Proposed Work . . . . .	154
6.3	Simulation 1: The Speaker Identification Accuracy (SIA) as a Function of the UBM Mixture Sizes {8, 16, 32, 64, 128, 256, 512 } for the I-vector Approach for 100, 200 and 400 Dimensions for the Original Speech Recordings of TIMIT Database . . . . .	155
6.4	Simulation 2: The Speaker Identification Accuracy (SIA) as a Function of the UBM Mixture Sizes {8, 16, 32, 64, 128, 256, 512 } for the I-vector Approach for 200, 400 and 800 Dimensions for Original Speech Recordings for the TIMIT Database . . . . .	156

6.5	Simulation 3: The SIA for the I-vector Approach for 100, 200 and 400 Dimensions Under AWGN at Different SNR Levels $\{0, 5, 10, 15, 20, 25, 30\}$ dB Without Handset at UBM Mixture Size 256 for the TIMIT Database . . . . .	157
6.6	Simulation 4: The SIA for Different Gaussian Mixture Components (GMCs) for the I-vector Approach for 100, 200 and 400 Dimensions Under Street Traffic NSN Without Handset at Different SNR Levels $\{0, 5, 10, 15, 20, 25, 30\}$ dB at UBM Mixture Size 256 for the TIMIT Database . . . . .	158
6.7	Simulation 5: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Bus Interior NSN Without Handset at Different SNR Levels $\{0, 5, 10, 15, 20, 25, 30\}$ dB at UBM Mixture Size 256 for the TIMIT Database . . . . .	159
6.8	Simulation 6: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Crowd Talking NSN Without Handset at Different SNR Levels $\{0, 5, 10, 15, 20, 25, 30\}$ dB at UBM Mixture Size 256 for the TIMIT Database . . . . .	160
6.9	Simulation 7: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under AWGN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels $\{0, 5, 10, 15, 20, 25, 30\}$ dB at UBM Mixture Size 256 for the TIMIT Database . . . . .	161
6.10	Simulation 8: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Street Traffic NSN with G.712 Type Handset at 16 kHz at Different SNR Levels $\{0, 5, 10, 15, 20, 25, 30\}$ dB at UBM Mixture Size 256 for the TIMIT Database . . . . .	162
6.11	Simulation 9: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Bus Interior NSN With G.712 Type Handset at 16 kHz at Different SNR Levels $\{0, 5, 10, 15, 20, 25, 30\}$ dB at UBM Mixture Size 256 for the TIMIT Database . . . . .	163
6.12	Simulation 10: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Crowd Talking NSN with G.712 Type Handset at 16 kHz at Different SNR Levels $\{0, 5, 10, 15, 20, 25, 30\}$ dB at UBM Mixture Size 256 for the TIMIT Database . . . . .	164

6.13	Simulation 1: The Speaker Identification Accuracy (SIA) as a Function of the UBM Mixture Sizes {8, 16, 32, 64, 128, 256, 512 } for the I-vector Approach for 100, 200 and 400 Dimensions for Original Speech Recordings (OSR) without Handset at UBM Mixture Size 256 for the SITW Database . . . . .	165
6.14	Simulation 2: The SIA for Different GMCs to the I-vector Approach for 100, 200 and 400 Dimensions Under AWGN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels {0, 5, 10, 15, 20, 25, 30} dB at UBM Mixture Size 256 for the SITW Database . . . .	166
6.15	Simulation 3: The SIA for Different GMCs to the I-vector Approach for 100, 200 and 400 Dimensions Under Street Traffic NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels {0, 5, 10, 15, 20, 25, 30} dB at UBM Mixture Size 256 for the SITW Database . .	167
6.16	Simulation 4: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Bus Interior NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels {0, 5, 10, 15, 20, 25, 30} dB at UBM Mixture Size 256 for the SITW Database . .	168
6.17	Simulation 5: The SIA for Different GMCs to the I-vector Approach for 100, 200 and 400 Dimensions Under Crowd Talking NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels {0, 5, 10, 15, 20, 25, 30} dB at UBM Mixture Size 256 for the SITW Database	169
6.18	Simulation 1: The Speaker Identification Accuracy as a Function of the UBM Mixture Sizes {8, 16, 32, 64, 128, 256, 512 } for the I-vector Approach for 100, 200 and 400 Dimensions for Original Speech Recordings (OSR) Without Handset at UBM Mixture Size 256 for the NIST 2008 Database . . . . .	170
6.19	Simulation 2: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under AWGN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels {0, 5, 10, 15, 20, 25, 30} dB at UBM Mixture Size 256 for the NIST 2008 Database . .	171
6.20	Simulation 3: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Street Traffic NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels {0, 5, 10, 15, 20, 25, 30} dB at UBM Mixture Size 256 for the NIST 2008 Database	172

6.21	Simulation 4: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Bus-Interior NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels {0, 5, 10, 15, 20, 25, 30} dB at UBM Mixture Size 256 for the NIST 2008 Database	173
6.22	Simulation 5: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Crowd Talking NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels {0, 5, 10, 15, 20, 25, 30} dB at UBM Mixture Size 256 for the NIST 2008 Database . . . . .	174
6.23	Simulation 1: The Speaker Identification Accuracy as a Function of the UBM Mixture Sizes {8, 16, 32, 64, 128, 256, 512 } for the NTIMIT Database . . . . .	175
7.1	Percentage Improvements for the I-vector Approach Compared with the GMM-UBM Approach for the TIMIT Database Under Different Environments . . . . .	181
7.2	Percentage Improvements for the I-vector Approach Compared with the GMM-UBM Approach for the SITW Database Under Different Environments . . . . .	182
7.3	Percentage Improvements for the I-vector Approach Compared with the GMM-UBM Approach for the NIST 2008 Database Under Different Environments . . . . .	183
7.4	The Best Feature With and Without Fusion I-vector Methods According to the Highest SIA for Three Databases . . . . .	184
7.5	Best SIA Performance for Feature Based Speaker Identification With and Without Fusion for Three Databases Using GMM-UBM Approach	185



# List of Acronyms

<i>ADT</i>	Audio Database Toolbox
<i>ANN</i>	Artificial Neural Network
<i>ANS</i>	Asymmetric Noise Suppression
<i>ASR</i>	Automatic Speech Recognition
<i>AWGN</i>	Additive White Gaussian Noise
<i>BFCC</i>	Bark Frequency Cepstral Coefficients
<i>BN</i>	Bottleneck
<i>BPNN</i>	Back Propagation Neural Network
<i>BWS</i>	Baum-Welch Statistics
<i>CDHMM</i>	Continuous Density Hidden Markov Model
<i>CDS</i>	Cosine Distance Scoring
<i>CFCCs</i>	Cochlear Filter Cepstral Coefficients
<i>CMN</i>	Cepstral Mean Normalization
<i>CMS</i>	Cepstral Mean Subtraction
<i>CMVN</i>	Cepstral Mean and Variance Normalization
<i>CMVNMFCCs</i>	Cepstral Mean and Variance Normalization Mel Frequency Cepstral Coefficients
<i>CMVNPNCs</i>	Cepstral Mean and Variance Normalization Power Normalized Cepstral Coefficients

## List of Acronyms

---

<i>CTIMIT</i>	Cellular-bandwidth Texas Instruments and Massachusetts Institute Technology
<i>DBN</i>	Dynamic Bayesian Networks
<i>DCF</i>	Detection Cost Function
<i>DCT</i>	Discrete Cosine Transform
<i>DET</i>	Detection Error Tradeoff
<i>DFT</i>	Discrete Fourier Transform
<i>DMFCC</i>	Dynamic Mel Frequency Cepstral Coefficient
<i>DNN</i>	Deep Neural Network
<i>DR</i>	Dialect Region
<i>DTW</i>	Dynamic Time Warping
<i>DWT</i>	Discrete Wavelet Transform
<i>EER</i>	Equal Error Rate
<i>ELM</i>	Extreme Learning Machine
<i>EM</i>	Expectation Maximization
<i>ERB</i>	Equivalent Rectangular Bandwidth
<i>FAR</i>	False Acceptance Ratio
<i>FD</i>	Feature Dimensions
<i>FDLP</i>	Frequency Domain Linear Prediction
<i>FFMTIMIT</i>	Far Field Microphone recording version of the of the Texas Instruments and Massachusetts Institute Technology
<i>FFT</i>	Fast Fourier Transform
<i>FIR</i>	Finite Impulse Response
<i>FRR</i>	False Rejection Ratio

## List of Acronyms

---

<i>FW</i>	Feature Warping
<i>FWMFCCs</i>	Feature Warping Mel Frequency Cepstral Coefficients
<i>FWPNCCs</i>	Feature Warping Power Normalized Cepstral Coefficients
<i>GFB</i>	Gammatone Filter Bank
<i>GFCCs</i>	Gammatone Frequency Cepstral Coefficients
<i>GFM</i>	Generalized Fuzzy Model
<i>GMCs</i>	Gaussian Mixture Components
<i>GMM</i>	Gaussian Mixture Model
<i>GMM – UBM</i>	Gaussian Mixture Model-Universal Background Model
<i>GRNN</i>	General Regressive Neural Network
<i>HMM</i>	Hidden Markov Model
<i>HTIMIT</i>	Handset TIMIT
<i>IDER</i>	IDentification Error Rate
<i>IMEL</i>	Inverse MEL
<i>IMFCC</i>	Inverse Mel Frequency Cepstral Coefficients
<i>JFA</i>	Joint Factor Analysis
<i>JNAS</i>	Japanese Newspaper Article Sentence
<i>LBG</i>	Linde-Buzo-Gray
<i>LDA</i>	Linear Discriminant Analysis
<i>LDC</i>	Linguistic Data Consortium
<i>LED</i>	Light-Emitting Diode
<i>LLR</i>	Log-Likelihood Ratio
<i>LPCC</i>	Linear Prediction Cepstral Coefficient
<i>LPRCC</i>	Linear Predictive Residual Cepstral Coefficient

## List of Acronyms

---

<i>LSH</i>	Locality Sensitive Hashing
<i>MAP</i>	Maximum A-Posteriori
<i>MFB</i>	MEL Filter Bank
<i>MFCCs</i>	Mel Frequency Cepstral Coefficients
<i>MIT</i>	Massachusetts Institute of Technology
<i>ML</i>	Maximum Likelihood
<i>MMFCCs</i>	Modified Mel Frequency Cepstral Coefficients
<i>MNN</i>	Multimodal Neural Network
<i>MSP</i>	Modulation Spectrum Processing
<i>NAP</i>	Nuisance Attribute Projection
<i>NFVQ</i>	Novel Fuzzy Vector Quantization
<i>NIST</i>	National Institute of Standards and Technology
<i>NSN</i>	Non Stationary Noise
<i>NTIMIT</i>	Network Texas Instruments and Massachusetts Institute Technology
<i>OSR</i>	Original Speech Recordings
<i>OSSRE</i>	Open Source Speech Recognition Engines
<i>PLDA</i>	Probabilistic Linear Discriminant Analysis
<i>PLP</i>	Perceptual Linear Prediction
<i>PNCC</i>	Power Normalized Cepstral Coefficients
<i>PNN</i>	Probabilistic Neural Network
<i>PRSIA</i>	Percentage Reduction in SIA
<i>RBF – NN</i>	Radial Based Function Neural Network
<i>ROC</i>	Receiver Operating Characteristics

## List of Acronyms

---

<i>RPLP</i>	Revised Perceptual Linear Prediction
<i>RR</i>	Recognition Rate
<i>RT</i>	Radon Transform
<i>RWCP</i>	Real World Computing Partnership
<i>SAD</i>	Speech Activity Detector
<i>SGGS</i>	Shri Guru Gobind Singhji
<i>SIA</i>	Speaker Identification Accuracy
<i>SISs</i>	Speaker Identification Systems
<i>SITW</i>	Speakers In The Wild
<i>SNRs</i>	Signal to Noise Ratios
<i>SRC</i>	Sparse Representation Classifier
<i>SRE</i>	Speaker Recognition Evaluation
<i>SSC</i>	Speech Separation Challenge
<i>STFT</i>	Short Time Fourier Transform
<i>SVM</i>	Support Vector Machine
<i>TI</i>	Texas Instruments
<i>TIMIT</i>	Texas Instruments and Massachusetts Institute Technology
<i>TISI</i>	Text-Independent Speaker Identification
<i>TVS</i>	Total Variability Space
<i>UBM</i>	Universal Background Model
<i>VB</i>	Variational Bayes
<i>VQ</i>	Vector Quantization
<i>WCCN</i>	Within Class Covariance Normalization
<i>WH</i>	With Handset
<i>WOH</i>	Without Handset

# Chapter 1

## Introduction

### 1.1 Background to Speaker Identification

Speaker recognition systems recognize individuals using their voice biometric. Such systems locate an individual's identity based upon who they are, rather than what they have or remember, such as an ID card or password. Speaker recognition itself is the operation of automatically recognizing who is speaking, depending on individual information contained in speech waves. This information is then compared to the entire biometric authentication across several biometric samples [2]. Speaker recognition can be achieved by identification and verification systems to verify an individual's purported identity from their voice. Verification systems are slightly different from speaker identification, which decides if a speaker is a particular person or is among a group of individuals [3]. With speaker identification, human speech from an individual is used to identify who is speaking. Fig. 1.1 provides a general block diagram of a speaker identification system. This system has two main stages, training and testing. Training, which is also called enrolment, involves the training process for all speakers who are to be identified, including the speech from each individual; testing processes are used to match the training and testing samples to the same speaker, as well as to recognize different speakers. Usually, the training phase is conducted off-line as a portion of the system formation and before the system is circulated. In testing, the true operation of the system is carried out, and the speech from an unidentified utterance is compared online against each of the trained speaker models [4]. Fig. 1.2 illustrates the speaker recognition tasks. Fig. 1.2-Part A illustrates how a

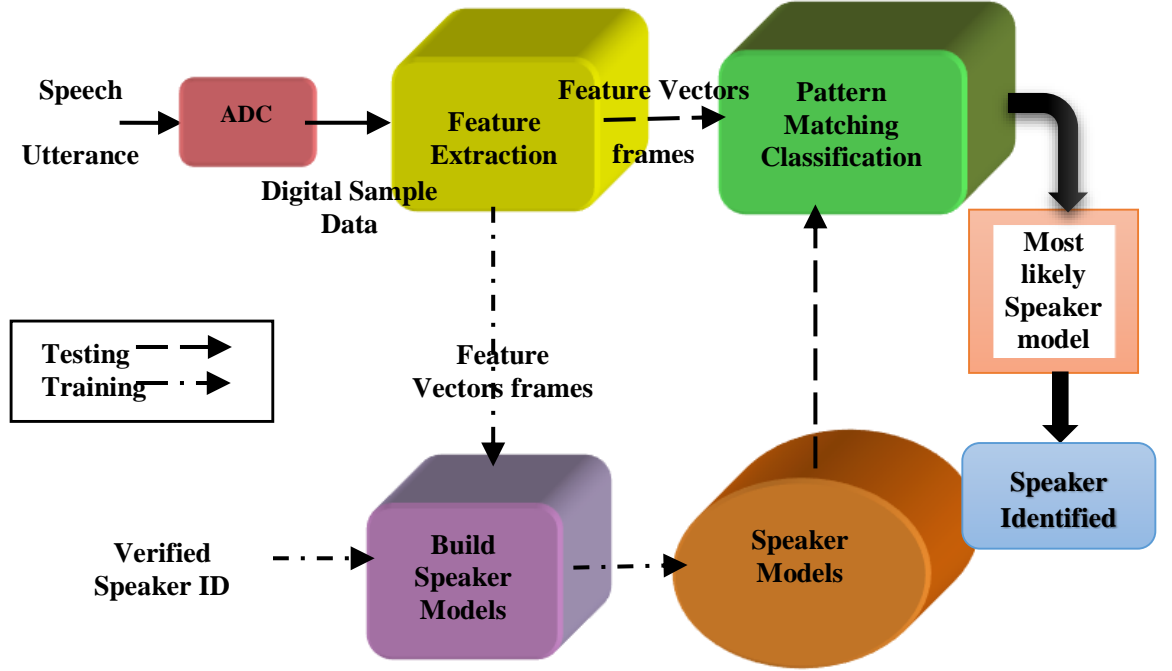


Figure 1.1: Block Diagram of a Speaker Identification System [3]

voice, a language or words can be recognised. A speaker identification system can be categorized into two types: closed and open set, and this is shown in Fig. 1.2 Parts B, C and D, which show closed set speaker identification, open set speaker identification, and speaker verification, respectively. Generally, speaker identification is one to N matching (where N is the number of speakers).

With closed set identification, the unknown individual belongs to a pre-existing pool or database of speakers (speaker models), and the next step is to match a speaker from the pool with the unknown speech [4]. Closed set identification is an exemplary independent community where the group members are known in terms of their speaker profiles, which are kept in a database. Identification is thus within this section and no users from outside are included within the model [4] [5], Fig. 1.2, Part B illustrates this task.

In open set identification, as depicted in Fig. 1.2 Part C, the unknown individual originates from the general population [4] [5]. Most speaker identification applications are open set, meaning that the unknown speaker may not be within the group of speaker models. In this situation, if no acceptable match is achieved, a no-match decision is provided [5]. Thus, the main aim of an open set identification system is to reveal whether the speaker belongs to the database of

## 1.1 Background to Speaker Identification

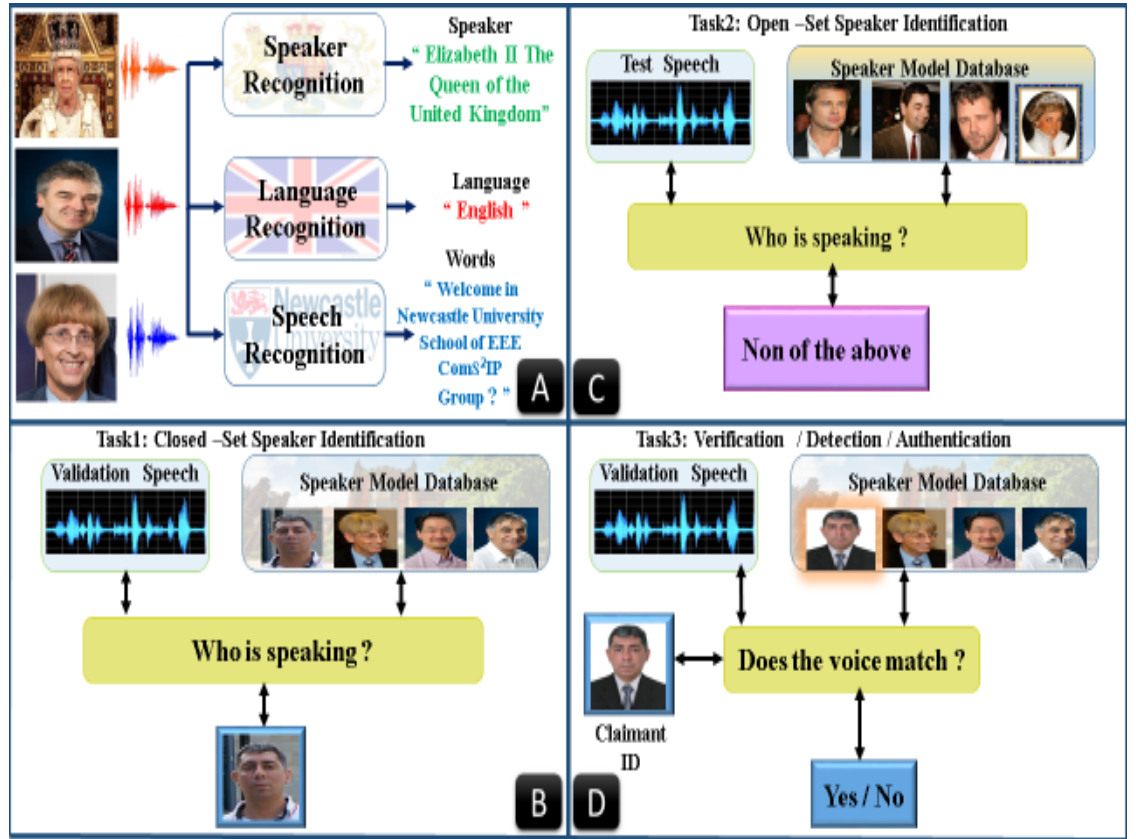


Figure 1.2: The Speaker Recognition Tasks: (A) Speaker, Language and Speech Recognition; (B) Task 1: Closed Set Speaker Identification; (C) Task 2: Open Set Speaker Identification; (D) Task 3: Verification, Detection and Authentication

unknown speakers, and the speaker is rejected if they do not belong to the pool [4] [5].

With speaker verification, human speech from an individual is used to verify the purported identity of that individual via one to one matching, as explained in Fig. 1.3 and in Fig. 1.2 Part D [3]. In this situation, the unknown speech sample is matched only with the speaker model whose label is identical to that of the identity being examined. If the value of the matching is suitable, the identity claim is accepted, but rejected if the claim is otherwise. For a one-speaker target group speaker verification is a particular case of open-set speaker identification. Only one comparison is required in a speaker verification trace; therefore, the performance of speaker verification is independent of speaker population size [4]. As in speaker identification, the initial formation of the system is executed through enrolment or training by verifying each speaker in the system, using samples of speech provided to train the model to the speaker. In testing, verification happens when the person



## 1.1 Background to Speaker Identification

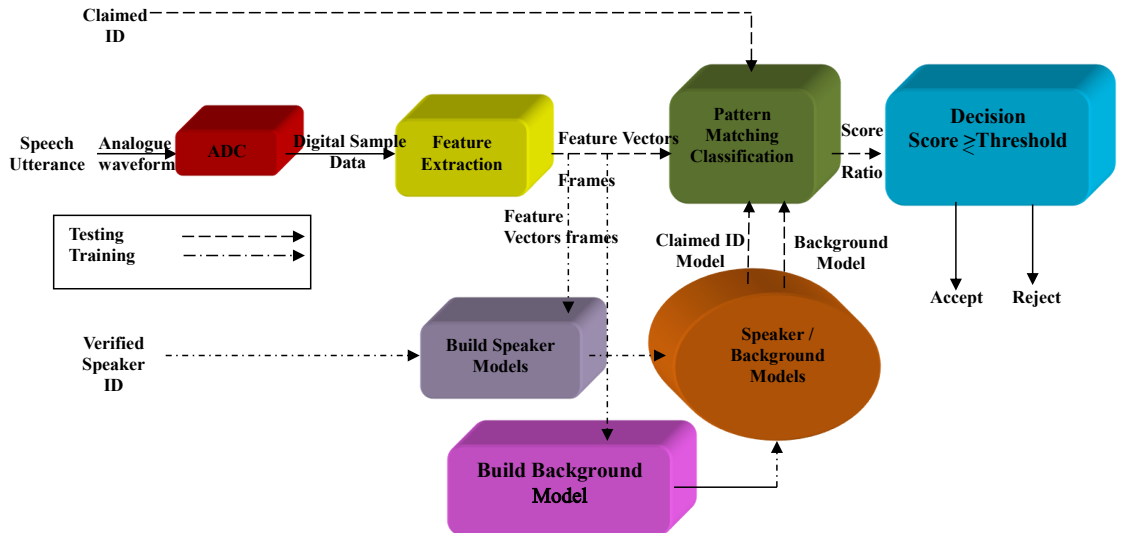


Figure 1.3: Block Diagram of a Speaker Verification System [3]

has to produce a claim as to who they are, and then the system verifies if the claim is true or false. The speech of an unknown individual in speaker verification is compared against each claimed identity and all other speakers or background models or the imposter. Then, the proportion of the two measures is taken and matched to a threshold; if the ratio is above the threshold, then the claim is accepted, but if the ratio is below the threshold then the claim is rejected as false [3].

In speaker detection, a third task of speaker recognition has recently been defined by the National Institute of Standards and Technology (NIST). In this task, an unknown speech sample is provided to determine whether one of a specific set of known speakers is present in the sample. When the unknown sample includes another speech from more than one speaker, then the task is more complicated. An example of this could be a telephone conversation between two people. In this situation, a further task called speaker tracking is possible; when the detected speaker is talking, it is essential to locate the inter values in the test sample [3]. One other application of speech detection concerns multi-speaker speech samples for speaker tracking, indexing and segmentation [4].

## 1.2 Aims

The main aim of this study is to investigate a new robust closed set speaker identification system using four feature combinations based on Mel Frequency Cepstral Coefficients (MFCC) and Power Normalized Cepstral Coefficients (PNCC). For each feature extraction method, two types of normalization are employed, Cepstral Mean and Variance Normalization (CMVN) and Feature Warping (FW). The aims can be summarized as follows:

- To use two different approaches to improve Speaker Identification Accuracy (SIA) with two different state of the art models using Gaussian Mixture Model-Universal Background Model (GMM-UBM) and I-vector, and to determine their effects in modelling speakers and identifying them. Fair comparisons will then be made between the two models.
- To study the effect of two classifier methods on the SIA, the traditional Log-Likelihood Ratio (LLR), and the Extreme Learning Machine (ELM).
- To test the system using 120 speakers from four databases (in total 480 speakers with 4,800 speech utterances): TIMIT Acoustic-Phonetic Continuous Speech Corpus, the Speakers In The Wild (SITW) Speaker Recognition Challenge, the 2008 NIST Speaker Recognition Evaluation Training Set Part 2-2011, and the NTIMIT telephone bandwidth version of TIMIT database (Network TIMIT).
- To study the quantitative perspective of noise and handset effects on SIA for two models, GMM-UBM and I-vector, by adding various background noises with handset when testing the speaker identification system through a wide range of Gaussian mixture components, namely: mixture sizes {8, 16, 32, 64, 128, 256, 512} in original speech recordings and seven ranges of Signal to Noise Ratio Levels (0-30 dB) with step size 5 dB for challenging environments, utilizing Additive White Gaussian Noise (AWGN), and different types of Non-Stationary Noise (NSN). These are street traffic, a bus interior, and crowd talk.
- To provide fair comparisons between different databases, and between different approaches of I-vector and GMM-UBM, in addition to creating

### 1.3 Motivation

---

multi-dimensional I-vectors, and then studying their effects on the SIA.

- This study provides benchmark evaluations of three databases for other researchers working in the speaker identification field.

## 1.3 Motivation

Although the GMM-UBM approach is well established, no previous study has comprehensively considered three databases, one of which only appeared in 2016, nor the effect of such a wide range of NSN and handset effects. However, many researchers have used the I-vector for speaker verification, and in this study a multi-dimensional fusion-based I-vector model are exploited for speaker identification with a thorough evaluation using four databases with a wide range of environmental noise conditions.

## 1.4 Objectives

The main objective in this thesis is to exploit different fusion techniques to improve the SIA. For this purpose, this study presents seven fusion methods: weights, maximum, mean, cumulative, interleaving, and two types of concatenated fusion I-vectors. In addition, the study applies both the Cepstral Mean and Variance Normalization (CMVN) and Feature Warping (FW) to mitigate noise and linear channel effects. Furthermore, the same settings are applied by using the same number of utterances, sampling rate, speaker numbers, and number of training and testing samples etc. Particular data from the SITW determines which is fusion-based.

## 1.5 Difficulties with a Speaker Identification System

Many factors can contribute to verification and identification errors. Some of these problems are caused by humans while others are environmental. These problems include [6] [7]:

## 1.5 Difficulties with a Speaker Identification System

---

- Human Comprehension: words generally come spontaneously even though there is a grammatical structure, and statistical methods have improved the prediction of words; however, there remains a problem with modelling world knowledge.
- Body language: human speakers use body gestures, such as waving hands, moving eyes etc. but in Automatic Speaker recognition, such paralinguistic features are not available.
- Noise: any unwanted information in a speech signal is called noise, for example a ringing clock, hearing another speaker talking at the same time, or a TV playing. Another type of noise is the echo effect, in which the speech signal rebounds on a surrounding object and as a result a few milliseconds later it appears in the microphone.
- Spoken language: this is not equal to written language as writing is more structured than spoken language. In addition, speaking is a two-way active communication, while writing is passively communicated. Disfluencies in speech, such as slips of the tongue, repetitions, and unexpected changes of subject, are also present in ordinary speech.
- Channel variability: one of the most critical difficulties faced in speaker recognition systems is channel mismatch. Anything that impacts on the acoustic waveform and its content to the speaker can also affect its discrete representation in the computer, just as different kinds of microphones and noise signals can change over time.
- Signal variation problems and time lapse effects may cause variation.
- Speaker variability: each person speaks differently to others.

Not only is each speaker's speech different, but there is also variation within any specific speaker. Such divergences can be [7]:

I) Speaking Style: the personality of each person is individual, causing all speakers to talk differently, at different times. In addition, humans express emotions by changing the behaviour of their speaking style, and this can be achieved in different ways by each speaker; therefore, speaking in a public area may be different from speaking with friends or teachers.

## 1.6 Speaker Recognition Applications

---

II) Realization: over time the speech realization has changed. When words are repeatedly pronounced, the resulting speech will never be the same, and so small differences in the acoustic wave are created.

III) Speaker Sex: males and females are different, as males have a longer vocal tract than females. The female pitch is almost twice that of males. The average basic frequency for adult males is approximately 100Hz, 200Hz for adult females, and 300 Hz for children.

IV) Dialects: dialects can be categorized as regional and social. Regional ones include features of vocabulary, grammar, and pronunciation, according to a geographical zone, while social dialects are determined by the social group that the speaker is in.

## 1.6 Speaker Recognition Applications

Speaker recognition has many applications for verification and identification tasks. Some are suitable for both tasks, and these are listed below [3], [5] and [6]:

- Telephone banking
- Telephone shopping
- Voice mail
- Control of access to services such as mobile banking, voice dialling, and mobile shopping
- Remote access to computers
- Database access services
- Information services
- Security control for a confidential information area
- Forensics
- Intelligent answering machines
- Remote credit card purchases
- Surveillance, monitoring and automated ID

# 1.7 Thesis structure

This thesis is organized as follows:

- Chapter One includes an overview of the speaker identification system, then the main aims and objectives. In addition, the chapter handled the difficulties and the applications for the speaker recognition task.
- Chapter Two includes the background and a related literature review in terms of the GMM-UBM and I-vector approaches for the speaker identification.
- Chapter Three focuses on different databases; four were used in this thesis: the TIMIT Acoustic-Phonetic Continuous Speech Corpus, the Speakers in the Wild (SITW) Speaker Recognition Challenge 2016, the 2008 NIST Speaker Recognition Evaluation Training Set Part 2, and the NTIMIT, which is a telephone bandwidth version (Network TIMIT). However, this chapter includes the family of TIMIT databases and some other databases which were not used in this thesis, and the reasons for excluding them. Furthermore, the chapter presents Speaker Identification Accuracy (SIA) and how the performance accuracy was measured in this thesis.
- Chapter Four presents the first contribution chapter, and includes speaker identification using the GMM-UBM approach with fusion and evaluated on original speech recordings. This chapter includes four main simulations using different fusion strategies for original speech recordings text independent speaker identification. These are: late fusion, early fusion, and early-late fusion. Late fusion is for concatenated static and dynamic features, and all the above scores are statistically independent normalized scores.
- Chapter Five shows the second contribution chapter using closed set speaker identification using the GMM-UBM approach with fusion for challenging environments with three databases.
- Chapter Six presents the third contribution chapter on fusion-based speaker identification using multi-dimensional I-vectors in challenging environments for four databases. This chapter answered three significant questions: the first question was how far the multi-dimensional I-vectors affect the SIA; the second question is how far the feature combination of I-vector, SNR level, UBM

## 1.7 Thesis structure

---

mixture sizes, with and without fusion, affected the SIA; the final question is whether the SIA is better with the I-vector or GMM-UBM methods of speaker identification.

- Chapter Seven is the summary of the thesis and the main results in terms of the databases, fusion techniques, speaker identification approaches. The thesis conclusions and possible suggestions for improving the SIA in future work are also presented here.

# Chapter 2

## Background and Related Literature Review

This chapter is focused on the background and literature review in terms of a new speaker identification system, as explained in Fig. 2.1. Modelling, fusion techniques and classification are explained in more depth in the contribution chapters (Chapters 4, 5 and 6). For example, the Gaussian Mixture Model (GMM) with a universal background model (UBM) is described in Chapters 4 and 5, and some fusion methods are also considered. The main concept of the I-vector approach and the mathematical model is fully explained in Chapter 6, where seven fusion methods are also considered. The background for classification methods such as the maximum likelihood and ELM are discussed in both Chapter 5 and 6. This chapter includes general concepts behind the speaker identification system, including the feature extraction and normalization methods used in this thesis. In addition, the background for various types of databases are presented in Chapter 3, which includes the description of the four databases used in this thesis. Different types of databases are also considered in Chapter 3. A literature review is provided for each contribution chapter, this chapter summarizes the literature in related work for other state of the art speaker identification systems.

### 2.1 The auditory perception system

Any simple communication system can be divided into the three main sections: transmitter, channel and receiver; similarly, human speech is comprised of the



## 2.1 The auditory perception system

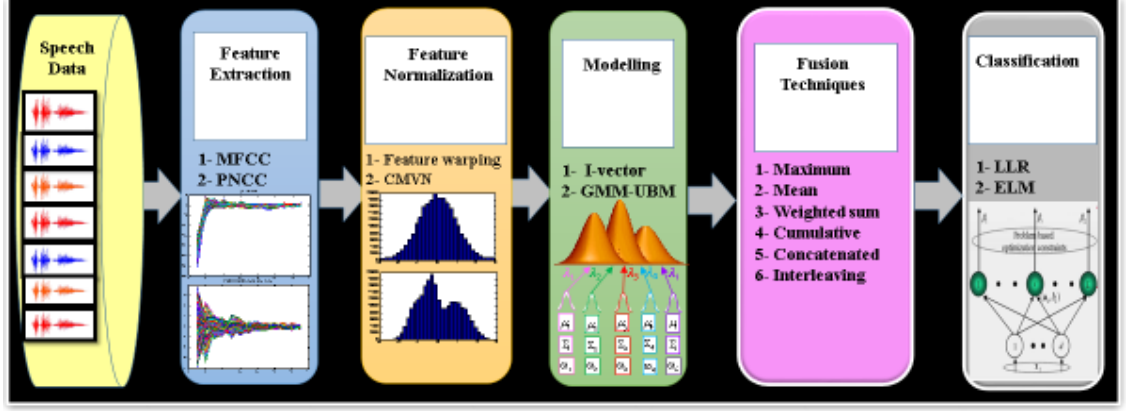


Figure 2.1: The Main Speaker Identification Scheme in This Thesis

process of transferring speech from the speaker (the transmitter) to the listener (the receiver) through free space (a noisy channel), as shown in Fig. 2.2 [8]. In addition, the vocal system is practically responsible for speech production, while physically the responsibility is returned to the brain; likewise, the auditory system is practically responsible for speech perception in the listener, whereas the auditory nerve system is physically responsible. The two major parts of the body responsible for auditory perception are the ears and brain; the peripheral auditory system represented by the ears handles received speech signals on the basilar membrane by converted them to a mechanical vibration paradigm. Thereby, successive pulses can be transmitted through the auditory nerve, and then the brain (the auditory nervous system) extracts the perceptual data. The human ear

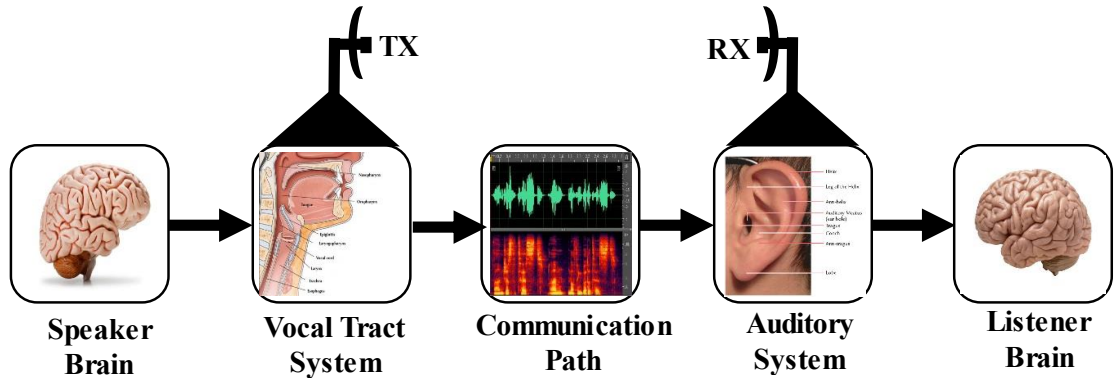


Figure 2.2: Human Communication System [8]

(the peripheral auditory system), as depicted in Fig. 2.3 is constructed of three

## 2.1 The auditory perception system

---

major parts: the outer, middle and inner ear, as shown in [9]. Fig. 2.4 was taken from the Manchester adult cochlear implant programme via Google. According to Fig. 2.4, the cochlea is the major part of the inner ear, in which the sound is represented to the brain through communication with the auditory nerve. This figure is also available on Google via [10]. Furthermore, human cochlea figures and cochlea implants can be viewed via by Wiley Online Library [11].

Moreover, the cochlea can be envisioned as a filter bank in which a frequency-to-location conversion is achieved, where the higher frequencies initiate a response in the filters closest to the cochlear base. In contrast, the lower frequencies cause a response in those filters closest to the cochlear apex. It can be seen through the cochlea's behaviour that the human hearing system has non-linear characteristics modelling generative speech. Fletcher researched the modelling of the natural response using frequency scales that are a good representation of the human perception system. Therefore, Fletcher's work (1940) recognised critical bands and pointed to the existence of the cochlear response. The Mel and Bark frequency scales were used for this purpose, which is common and widely used in both speech and speaker recognition. In this chapter, non-linear behaviour is explored with Mel Frequency Cepstral Coefficients

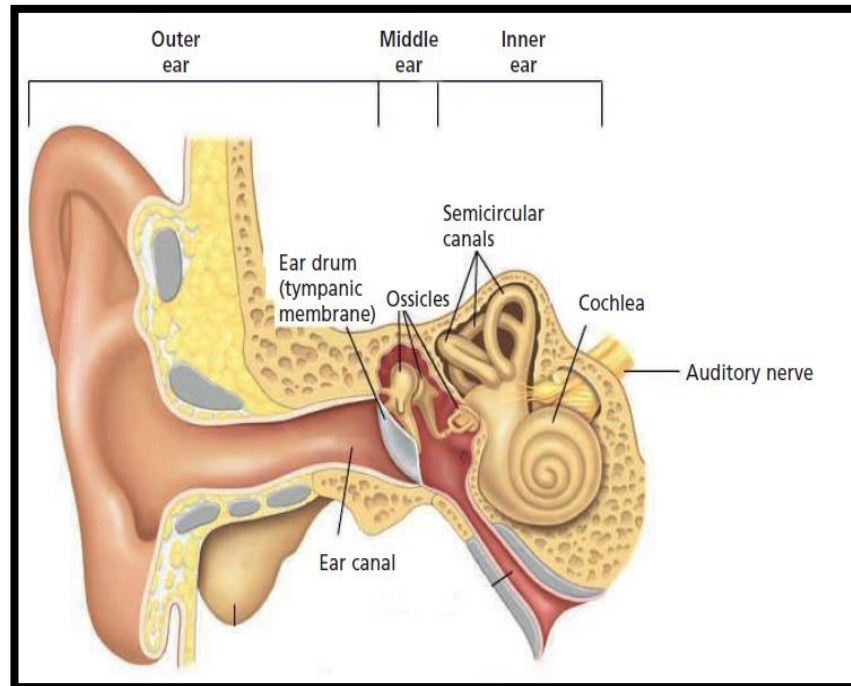


Figure 2.3: A Cross Section of the Human Ear [9]

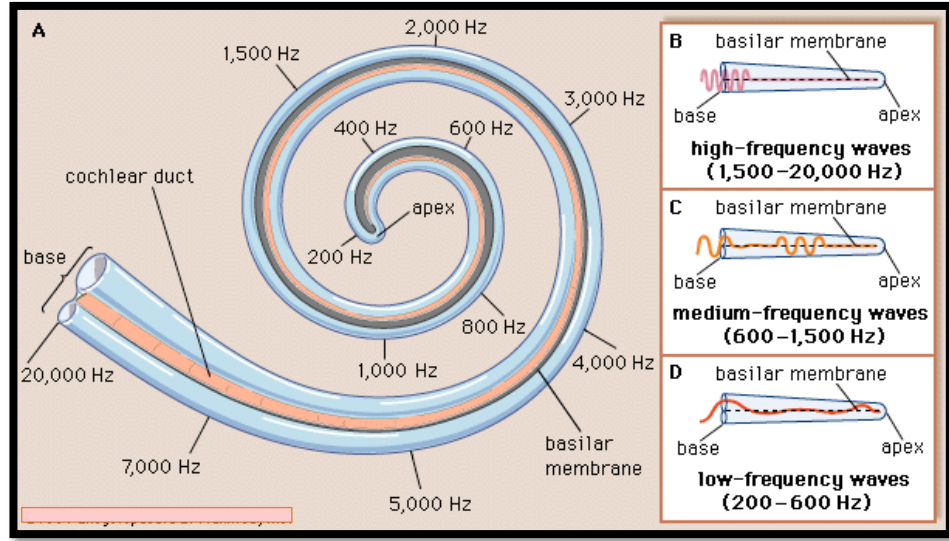


Figure 2.4: Implant Cochlea [10]

(MFCC) features to mimic the human ear. However, as an alternative to the logarithmic non-linearity produced by MFCC features, Power law non-linearity can be achieved by using Power Normalized Cepstral Coefficients (PNCCs) features as a feature extraction method. Hence, the Mel Filter Bank can be replaced by Gammatone Filter Bank (GFB).

## 2.2 Background to Speaker Identification Systems

This section describes the essential parts of a speaker identification system, including six major stages used to implement this system. These stages are: feature extraction, feature normalization, speaker modelling, classification, fusion techniques, and calculating SIA. The first part focuses on feature extraction, by which the speech signal is transformed into a compressed form with effective representation [6]. With speaker modelling, the feature vectors are trained via a statistical model of the speaker's acoustic space. For classification (matching), scoring is undertaken between the testing processes of feature vectors (unknown speakers) and the modelling process of speaker models (the known speaker(s)) to determine the match score.

### 2.2.1 Feature Extraction

Feature extraction is the transformation of the raw speech signal into the data set with a reduced number of variables covering the most significant information this process takes into the consideration. This may be on the principle of cost, or the need to remove unwanted information, such as redundancy, or to reduce complexity from the classifier so as to acquire a better performance [6]. In feature extraction, the feature dimensionality can be reduced by using all variables and the data are converted (using non-linear or a linear transformation). Thus, the goal is to replace the original variables with a smaller set of underlying variables. There are many reasons for executing feature extraction [6]: 1) to reduce the input data bandwidth, with resulting improvements in speed and reduced data requirements; 2) to enhance performance by providing an appropriate group of features for simple classifiers; 3) to remove or reduce unwanted information from the speech signal, such as redundancy; and 4) to recover features or new significant implicit variables so that the data may be easily observed and the structure of, and relationships in, the data identified. This chapter focuses on two major features, namely MFCC and PNCC features, and both are considered for all contribution chapters in this thesis.

### 2.2.2 Mel Frequency Cepstral Coefficients (MFCCs)

Cepstrum is the inverse Fourier Transform of the log-spectrum. Bogert et al. (1963) invented the word Cepstrum, which comes from reversing the letters in the first syllable of the word spectrum. Fig. 2.5 shows the main block diagram for MFCC features, classified into five sections [12] [13]:

- a. Pre-emphasis
- b. Frame blocking and windowing
- c. Fast Fourier Transform
- d. Mel-scaled filter bank
- e. Cepstrum

The main aim of the Pre-emphasis stage is to compensate for the high frequency part of the speech signal that was suppressed during the human sound production

## 2.2 Background to Speaker Identification Systems

---

mechanism. A first order Finite Impulse Response (FIR) high-pass filter is used to achieve pre-emphasis filtering [3] [13]. This thesis has used pre-emphasis parameter of 0.96 to mirror the work in [1]. In framing and windowing Part (b), the speech signals are non-stationary, and each speech signal can be divided into frames which are then analysed independently, as a feature vector has stationary behaviour. These frames maintain a length which avoids degradation in the frequency resolution (when too short), but can also capture the local spectral properties. In addition, to reduce the discontinuities of the speech signal at the edges of each frame, a tapered window is applied to each one. A Hamming window is the most common type and was used in this thesis at 16 ms frame length and 8 ms intervals as well as the sampling frequency is 16 kHz as in [1]. The Hamming windowing was employed to avoid and reduce any unnatural discontinuities at the edges of each frame of the speech signal [12] [13]. Part (c) is the Fast Fourier Transform (FFT), an N-point FFT is employed. Uniform space with  $\frac{N}{2}$  values of complex spectrum are produced between 0 to  $\frac{F_s}{2}$  ( $F_s$  is the sampling frequency 16 kHz). The magnitude of FFT is only used in speech processing by ignoring the phase information [3]. The magnitude coefficients of the N-point FFT are converted to the triangular filter bank with K values ( $K = 40$ ). The cross wise multiplication between the N-point FFT with the weighting function is the K filter bank. Then accumulating the results and denoting the output by  $i$ th filter bank  $Y(i)$ . In Part (d), a Mel-scaled filter bank is employed, which has a triangular frequency response. The behaviour of the Mel-scale is linear frequency spacing below 1 kHz (see equation 2.1). However, above 1 kHz, this behaviour is logarithmic spacing and the bandwidth and spacing are calculated by a constant interval of Mel-frequency [13]:

$$Mel(f) = \begin{cases} f & \text{if } f \leq 1kHz \\ 2595 \log_{10} \left(1 + \frac{f}{700}\right) & \text{if } f > 1kHz \end{cases} \quad (2.1)$$

With the speech signal, the information obtained by the low frequency components is more important than that carried by high frequency components. Therefore, Mel filter banks have non-uniform frequency spacing to emphasise the low frequency components. For this reason, the filter bank has more filters in low frequency zones compared to high [12] [13]. Part (e), to convert the cepstrum to the time domain, the log of the Mel spectrum, has to be transformed back to time in the final step to

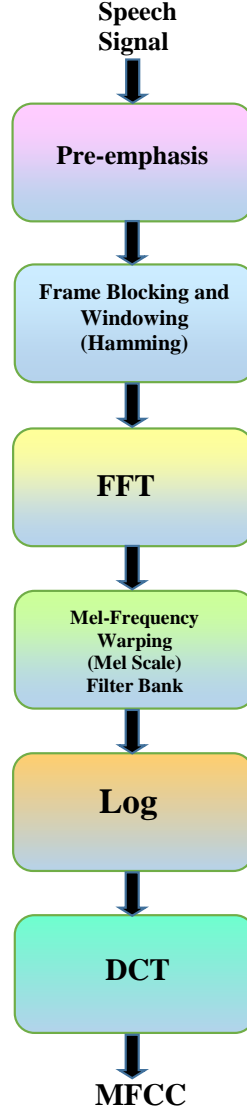


Figure 2.5: Mel Frequency Cepstral Coefficients Features (MFCCs) [3] [13]

produce the Mel Frequency Cepstrum Coefficients (MFCCs). Both the Mel spectrum coefficients and their logarithms are real numbers, so the cepstral features are a good representation of the local spectral properties. Then, the Discrete Fourier Transform (DCT) is used to convert the log of filter bank spectral values to the  $L$  cepstral coefficients [3] [13]. The MFCC is determined via the following equation [13]:

$$C_n = \sum_{i=1}^K \left( \log_{10} Y(i) \cos \left[ i \frac{2\pi n}{N} \right] \right) \quad (2.2)$$

where:  $C_n$  is the Cepstrum coefficients,  $n = 1, 2, \dots, L$ ,  $N$  is the number of FFT points ( $N = 512$ ),  $K$  is the number of channel filter banks ( $K = 40$ ), and  $Y(i)$  is the output of the  $i$ th filter bank.

### 2.2.3 Power Normalized Cepstral Coefficients (PNCCs)

According to Fig. 2.6, there are three major stages in computing the PNCC feature extraction algorithm [14], [15] and [16]:

1. Initial processing
2. Environmental processing
3. Final processing

In the initial processing, several stages are included and they are: pre-emphasis, in which the speech signal pre-emphasis parameter is set at 0.96 to mirror [1], and then the Short Time Fourier Transform (STFT) is utilized using a Hamming window of 16 ms frame length and 8 ms overlap interval between the frames. Then, the outputs to the magnitude of STFT are squared. In addition, 40 channels of Gammatone Filter Bank are employed to cover the telephone speech bandwidth(300-3,400) Hz. The centre frequencies of the Gammatone filters are linearly spaced in the Equivalent Rectangular Bandwidth (ERB), which is the auditory frequency scale. In speech recognition, the PNCC gives better accuracy than the MFCC in the presence of white noise [15], and this property was exploited for the the speaker identification task in this thesis. For the environmental processing, two sub-stages are included: temporal processing and spectral smoothing, where both have a substantial impact on the accuracy performance in white noise. In addition, the estimation and compensation of noise are considered by using the medium time power. Also, asymmetric noise suppression is used to reduce the noise effect by spectral subtraction of the noise level, estimated from the power of non-speech segments. However, the final processing contains the DCT and then the Mean Normalization, as explained in equation (2.3).

$$\mu[m] = \lambda_{\mu} \mu(m-1) + \frac{(1-\lambda_{\mu})}{L} \sum_{l=0}^{L-1} T[m, l] \quad (2.3)$$

where:  $m$  represents the frame incident,  $l$  is the channel incident,  $L$  is the number of frequency channels, and  $\lambda_{\mu}$  is the forgetting factor which is equal 0.999, whereas the  $T[m, l]$  represents the time frequency normalization. The power-law non-linearity is produced by a supposed value (1/15) that gives acceptable accuracy in white noise and without any significant impact on recognition accuracy in clean speech, as in

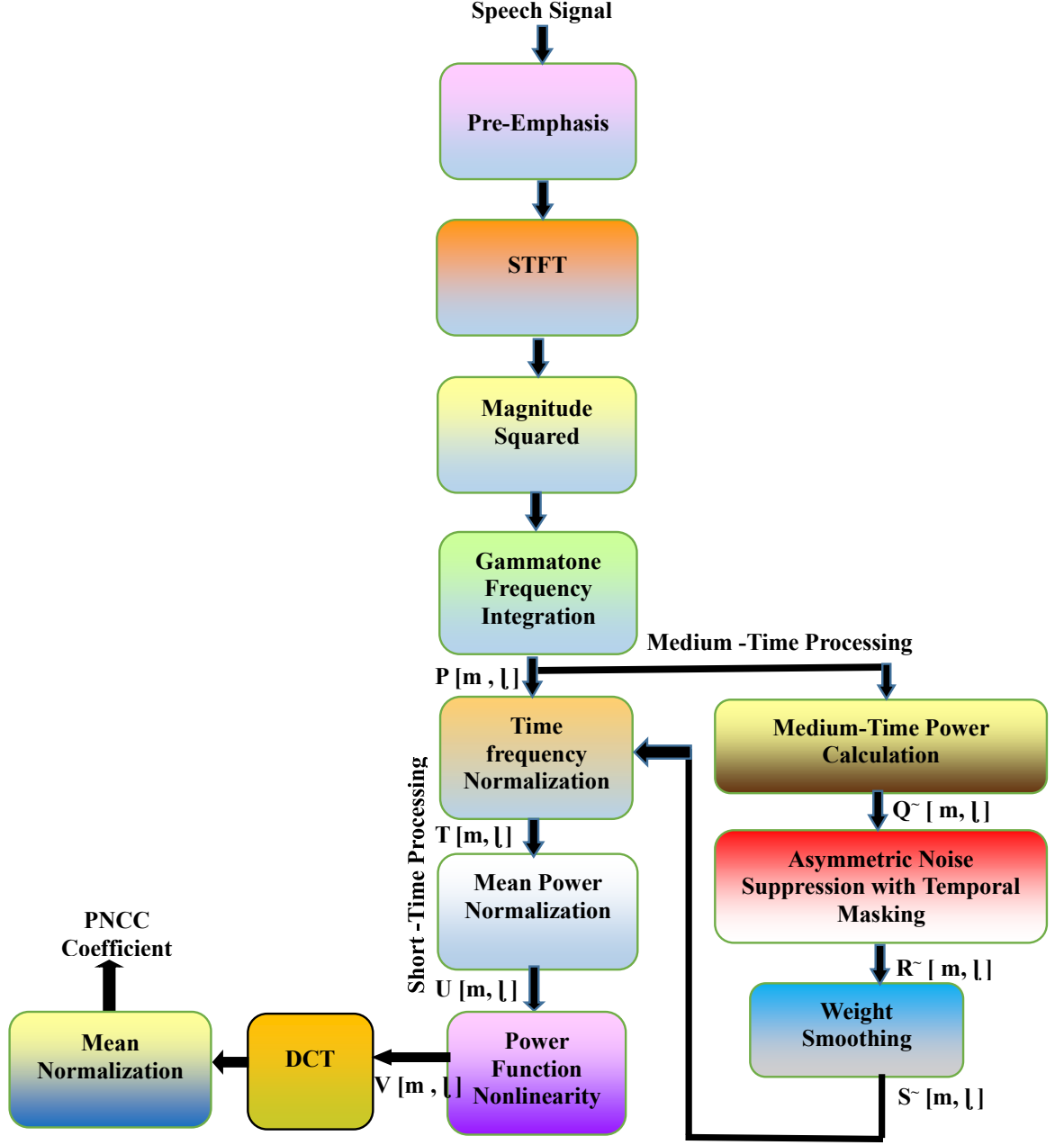


Figure 2.6: Power Normalized Cepstral Coefficients (PNCCs) Features [15]

equation (2.4), as explained in [14], [15] and [16].

$$V[m, l] = U[m, l]^{\frac{1}{15}} \quad (2.4)$$

where  $U[m, l]$ : is the normalized power.



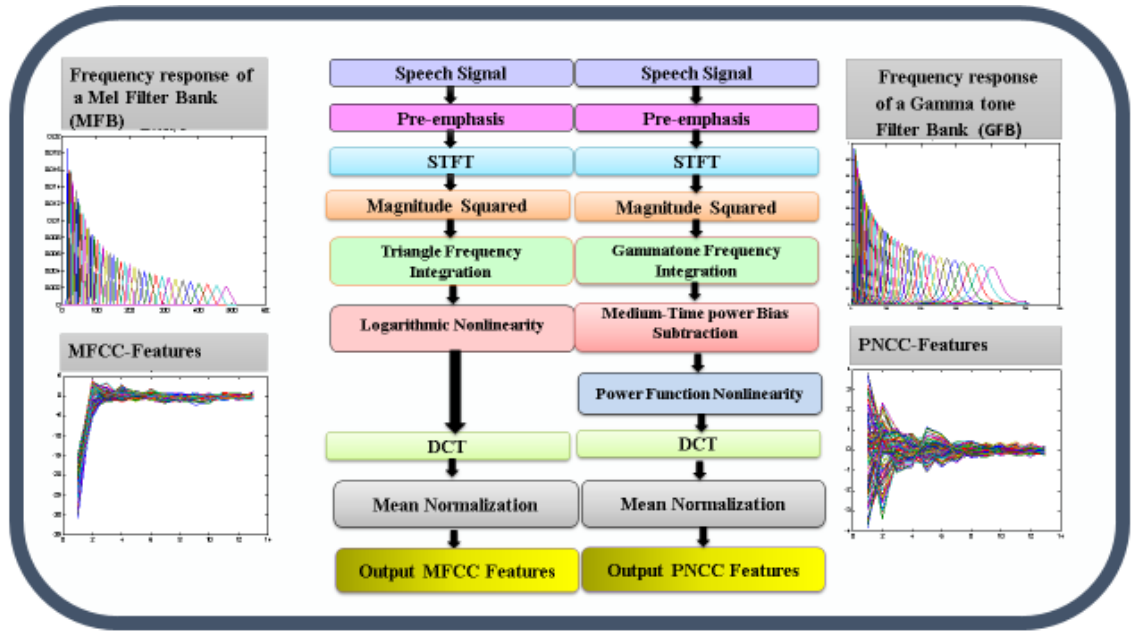


Figure 2.7: Comparison Between MFCC and PNCC Features Structure [6]

### 2.2.4 Comparisons Between MFCC and PNCC Features

Fig. 2.7 compares the structure of MFCC and PNCC features [6]. This figure depicts the main differences between the MFCC and PNCC features, which can be summarized by the following bullet points [14]:

- **MFCC-Features**
- Triangular / MEL Filter Bank (MFB)
- Logarithmic non-linearity
- Slightly less accurate in Automatic Speech Recognition (ASR) in presence of white noise
- The complexity has less computation than the PNCC as in [15]
- **PNCC-Features**
- Gammatone Filter Bank (GFB)
- Power law non-linearity
- Slightly better accuracy in ASR in presence of white noise
- The complexity has 33% more computation than the MFCC [15]

### 2.2.5 Feature Normalization Methods

Feature Compensation (Normalization) is common, and has been widely and effectively used for speaker recognition in both verification and identification tasks. Substantial aims can be achieved from using normalization, such as reducing the channel, handset transducer, and linear and non-linear channel effects. Numerous methods have previously been used for compensation features: Cepstral Mean Subtraction (CMS), Modulation Spectrum Processing (MSP), Short-term windowed and variance normalization CMS, CMVN and FW [17] [18] [19]. In this chapter, FW and CMVN are robust to additive noise and handset effects, as well as to mitigate the linear and non-linear channel effects [17]. This gives improvements and robustness to the speaker identification accuracy system [3]. The important aspects to compare between feature warping and CMVN are as listed below [20]:

- **Feature Warping**
  - **Purpose:** Gaussianization for the short-term over a sliding window
  - Over the specified time interval, and based on its rank in the array of sorted feature values, the middle frame in the window is normalised
  - Overall distribution of the feature stream is warped to the standard normal distribution
  - **Aim:** Mitigate the linear channel effects
- **CMVN**
  - **Purpose:** Over a sliding window, Cepstral mean and variance normalization
  - Over the specified time interval, based on the mean and variance computed, the middle frame in the window is normalised
  - Feature stream distribution is almost mapped to the standard normal distribution
  - **Aim:** Remove the linear channel effects

Other speaker identification parts: such as modelling, fusion techniques and classification, are discussed in more depth in the contribution chapters.

## 2.3 Literature Review for Speaker Identification

According to the main speaker identification scheme, this section can be divided into six subsections of related work: feature extraction, I-vector extraction, modelling, noise robustness and challenging environments, classification, and fusion technologies.

### 2.3.1 Feature Extraction

In 2005, Reda and Aoued [21] investigated artificial neural networks and the employment of MFCC for speaker recognition. The aim of using MFCC features was to imitate the human ear. The paper described two systems; System 1 was trained by an artificial neural network, whereas System 2 was trained using a vector quantizer design with Linde-Buzo-Gray (LBG) algorithm. To evaluate and compare the performance of the two systems with speech from different types of noise and sessions, 142 self collected subjects were employed. 203 subjects from CMU dataset and 60 from the ASR database were also selected. The testing recognition rates attained for System 1 were, respectively, 70.91%, 78.57% and 90.66% for ASR, self collected and CMU databases, respectively. However, System 2 obtained 80.01%, 85.71% and 69.33% for the ASR, self-collected and CMU databases, respectively. The effect of noise and handset were not considered in this paper. In addition, alternative features can be proposed and are more robust for noisy speech, such as the PNCC features considered in this thesis.

In 2010, Wang et al. [22] presented the integration between the phase information and MFCC features under noisy environments in terms of speaker recognition. The paper tested 35 speakers from the NTT database and 270 speakers from JNAS database. This paper has some drawbacks related to the use of Japanese speakers, which can not be fairly compared with standard English speakers used in other related work. In addition, using different settings, feature dimensions, number of speakers, and DFT samples makes any comparisons unfair. For example, 12 MFCC features from the NTT database compared with 25 in the JNAS database were then applied to the 25 ms window size. There were 512 samples of DFT and 10 ms for overlap, and for phase information the window size was 12.5 ms with 256 samples of DFT. A fair comparison between the databases from the first trend, and between the two features from the second trend, cannot therefore be made.

### 2.3 Literature Review for Speaker Identification

---

In 2010, Yujin et al. [23] proposed merging between the MFCC and Linear Prediction Cepstrum Coefficients for speaker recognition. To identify the speakers, the system employs the Dynamic Time Warping (DTW) and Vector Quantization (VQ). This paper showed that the highest recognition rate was achieved by integrating both features, with 40 speakers being evaluated using continuous digit speech (0-9). The system used LPCC coefficients with 12 order and 16 for MFCC features. The highest Recognition Rate (RR) from the combination was 97.12%. The main drawback of this paper was the use of different feature dimensions with different numbers of speakers to improve the RR. This is technically an error which will cause a combination of different features order (MFCC with the LPCC) to achieve a higher RR, and thereby gives unfair comparisons. In addition, the population size is limited.

In 2011, Ajmera et al. [24] developed a new feature extraction method based on the speech spectrogram by employing the Radon and the Discrete Cosine Transform (DCT) for text independent speaker identification. Basically, this paper converted the speaker recognition problem to pattern image recognition, and then machine learning tools were employed to resolve it. From the spectrogram to the speech signal, the acoustic features can be derived using Radon techniques. Hence, the speaker's voice pattern was obtained by calculating the Radon projections in various directions. However, to minimize the dimension to the feature vector, the DCT was applied to the Radon projection in order to achieve effective speaker features. The system was evaluated by using the Texas Instruments and Massachusetts Institute of Technology (TIMIT) and Shri Guru Gobind Singhji (SGGS) databases with 630 and 151 speakers, respectively. The best recognition rate achieved in this paper was 96.69% and 98.41% for the TIMIT and SGGS databases. The study omitted testing the system with a realistic database (such as the NIST) and different non-stationary noise types with handset were not considered. In addition, no fair comparisons in terms of number of speakers for the two databases can be made.

In 2012, Tazi et al. [25] developed a combination between Gammatone Frequency Cepstral Coefficients (GFCC) with the Cepstral Mean Normalization (CMN) and compared the MFCC baseline approach for a text independent speaker identification system. The system was conducted with a self-built database of 51 Arabic speakers (16 female and 35 male), with one utterance per speaker achieved for both the training (20 seconds) and testing (10 seconds). The feature vector dimension was

### 2.3 Literature Review for Speaker Identification

---

36 for both MFCC and GFCC features, with a window size of 32 ms and step size of 16 ms; the sampling rate was 16 kHz, and GMM mixture size was 8. In addition, testing under white noise was also considered with SNR levels from (0-40) dB. The GFCC with CMN approach attained a higher identification rate compared with MFCC in presence of white noise, at about 5.38%, when SNR levels were changed from 0 dB to 40 dB. In this paper, a realistic database and the effect of different non-stationary noise types and handset were missing. Also, fusion techniques might have been used to merge the two features to improve identification performance. Furthermore, using the SNR levels with (35 and 40) dB was worthwhile, but the effects were similar to the clean speech.

In 2012, Sumithra and Devika [26] developed five feature extraction techniques: the Linear Prediction Coefficient Cepstrum (LPCC), Revised Perceptual Linear Prediction (RPLP), Bark Frequency Cepstral Coefficients (BFCC), MFCC and Modified MFCC (MMFCC). These were used to study text independent speaker identification. Depending on the calculating time, the performance of these features was carried out and compared. Then, a Vector Quantization (VQ) codebook was utilized for speaker modelling. The system was tested by 100 speakers from the TIMIT database and windowing by 512 samples and the overlapping was 100 samples. The sampling frequency was 8 kHz with 40 filter banks for the BFCC, MFCC and MMFCC. MFCC attained the highest identification accuracy of 99.87% with a minimum distance of 4 and an initial centroid of 128. Challenging environments, such as the handset and stationary and non stationary background noise, and a realistic database such as NIST database, were not considered. Moreover, fusion based techniques could have been suggested to improve the performance accuracy.

In 2012, Trabelsi and Ayed [27] presented various feature extraction and normalization techniques for text independent speaker identification. These features extraction approaches were MFCC, LPC and Perceptual Linear Prediction (PLP), used with two normalization methods CMN, and rasta filtering. Then, a hybrid was created of the GMM with linear/non-linear Support Vector Machine (SVM) system. The paper showed the performance accuracy based on each feature extraction type and kernel functions. The system was conducted with 14 female speakers selected from DR1 to the TIMIT database. The hybrid GMM-SVM system showed that LPC and MFCC have better performance than PLP.

### 2.3 Literature Review for Speaker Identification

---

Furthermore, the best accuracy was obtained from delta and delta delta was achieved by LPC. In addition, normalization methods did not affect the identification accuracy as there is no session variability in the TIMIT database. The main drawback of this paper was that it did not consider any realistic database or study noise and handset effects. In addition, fusion methods could have been proposed for the feature extraction methods to improve the identification accuracy of the system.

In 2012, Ambikairajah et al. [16] presented PNCC features as an alternative to MFCC for robust speech recognition; hence, this paper presented it for speaker verification. In addition, the system was modelling by the I-vector approach and speakers were classified by Sparse Representation Classifier (SRC). In addition, the fusion score was compared for PNCC/I-vector and MFCC/I-vector approaches to improve the performance. The system was conducted using the NIST 2010 SRE, and showed that the best performance achieved 0.498 for DCF when fused with the SRC and both MFCC and PNCC features. However, Cosine Distance Scoring (CDS) was also considered as well as fusion based using CDS between the MFCC and the PNCC features; the best performance attained was 3.55% EER. The setting for this paper was to use 16 features for both MFCC and PNCC, and a Hamming window with 20 ms and 10 ms for overlap. In addition, the system was evaluated by the NIST 2010 database. In this thesis, the idea for using an I-vector based PNCC and MFCC was exploited for speaker identification by applying different feature combinations and classifiers. Also, this thesis modified the system to be more robust against AWGN and different non-stationary noise types with a handset effect using different databases. Furthermore, various I-vector fusion based methods can be used to modify the identification rate.

In 2013, Nidhyananthan et al. [28] developed the pitch based Dynamic MFCC (DMFCC) and MFCC features as well the integrated DMFCC and MFCC features with 19 coefficients for text independent language and speaker recognition. GMM was also used to model speakers. The system was tested with 120 self collected Tamil and English speakers recorded by GoldWave software employing a condenser microphone with 16 kHz at 16 bit mono. The performance measure used the IDentification Error Rate (IDER). The paper attained 5.8%, 2.9% and 1.2% for MFCC, DMFCC and the combination of both features, respectively. However, a realistic and standard database was missing from the paper, and noise and handset

### 2.3 Literature Review for Speaker Identification

---

effects were also not considered. In addition using the GMM approach has a drawback related to the limited Gaussian mixture components, and hence this affects identification accuracy. Furthermore, using error rate performance rather than identification accuracy was reported in this paper.

In 2014, Moinuddin and Kanthi [29] presented the GFCC to identify speakers under noise conditions, but the ability of the human ear was the main motivation for this paper. 23 feature dimensions from GFCC were compared with 13 from MFCC, and then the system was modelled with GMM. In addition, the modified MFCC (MMFCC) and GFCC (MGFCC) features were used, and instead of using the log, the cubic root was employed. The system was modelled by GMM and conducted using 630 speakers from both NTIMIT and TIMIT databases. The system showed that GFCC features give better identification accuracy compared with the traditional MFCC features in both clean (TIMIT) and noisy environments (NTIMIT). The drawback with this system was again that the effects of noise and handset, as well as realistic noise, were not considered. Moreover, fusion methods were missing in this paper, and these might have improved the accuracy.

In 2015, Almaadeed et al. [30] developed wavelet analysis for the feature extraction method, and this was used with multimodal neural networks for a speaker identification system. MFCC, wavelet packet transform, discrete wavelet transform, and wavelet sub-band coding were used. The system was tested with 34 speakers from the GRID database. The system used wavelet analysis and showed an improvement of 15% compared with the system with traditional MFCC. However, this work lacked speaker numbers and did not consider the effect of noise and handset, nor employ a realistic database.

In 2015, Zhang et al. [31] presented bottleneck feature mapping based on a Deep Neural Network (DNN) under dereverberation conditions for speaker identification of distant talking. This system used the Japanese Newspaper Article Sentence (JNAS) database for clean speech. However, to generate simulations for the dereverberation data various impulse responses were convoluted with clean speech. The Real World Computing Partnership (RWCP) database was selected, with eight multichannel impulse responses. In addition, the CENSREC-4 database was used to produce the artificial reverberant speech. However, again the paper did not study the effects of noise and handset, or consider a realistic database.

### 2.3.2 Literature Review in Terms of I-vector Extraction

This aspect is covered in depth in the contribution chapter, Chapter Six. In this section, only the important initial research on the I-vector for speaker verification is studied.

In 2005, Kenny [32] proposed the theory and algorithms of Joint Factor Analysis (JFA) for both sessions and speaker variabilities.

In 2007, Matrouf et al. [33] efficiently exploited the Factor Analysis Model for text independent speaker verification to tackle the session variability problem. Eigen channel MAP and various compensation methods were employed, such as GMM likelihood. In addition, Kernel based SVM was also applied. The system was evaluated by the NIST SRE 2005 and 2006, and measured by the EER and DCFmin. It attained a 50% improvement compared with the baseline of GMM-UBM.

In 2008, Kenny et al. [34] studied inter-speaker variabilities for speaker verification and the system was tested using the NIST 2006, and it attained a 10-15% reduction in EER. The JFA was proposed for this work and fusion for multiple systems was also considered as well as factor analysis with 200 channel factors and 300 speaker factors with less than 3% EER.

In 2008, Senoussaoui et al. [35] presented an I-vector extractor for both telephone and microphone speech with channel compensation methods using Linear Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) for speaker verification. The system employed two classifiers, the SVM and CDS. The system was tested by the NIST 2008 with interview data and the best performance was achieved by fusing the JFA with the CDS and SVM.

In 2012, Kenny [36] presented the iterative method using the Variational Bayes (VB) algorithm to reduce the running and training times for the extraction of the I-vector. The aim of this paper was to compare the VB based I-vector with the JFA to achieve an improvement in high dimensional I-vectors with the speaker verification. The paper also presented the accuracy for different I-vector dimensions (400, 800, 1200 and 1600) and the evaluation was carried out by the NIST 2010.

In 2015, Verma and Das [37] provided a survey of I-vector applications in speech processing. In this paper, the main concepts for JFA and the I-vector extractor were described, and various applications were presented such as speech diarization, accent,



## 2.3 Literature Review for Speaker Identification

---

dialect, speaker, emotion, and language recognition, as well as acoustic detection. In addition, hybrid methods were also considered such as the prosodic, combination of prosodic and cepstral approaches, and the phonotactic approach. Toolkits were also included, such as the MSR Identity Toolbox, LIUM speaker diarization, ALIZE 3.0 and The Kaldi speech recognition toolkit.

### 2.3.3 Modelling

This subsection can be subdivided into two main parts: different state of the art speaker identification modelling methods, and I-vector speaker identification.

Various state of the art speaker identification modelling methods have been used and can be categorized into three main sections according to the modelling method: Speaker Identification Systems (SISs) using the GMM model, and SISs using the GMM-UBM model, and SISs using various modelling methods.

In 1995, Reynolds [38] and likewise in [39] presented the GMM to represent speakers for both speaker identification and verification applications. The aim was to show the relationship between performance accuracy and population size for both clean and telephone speech. In addition, verification experiments were also considered. The identification system was conducted by TIMIT, NTIMIT and Switchboard databases, and for the verification system, YOHO database was also added for the evaluation. The Gaussian mixture components were limited by 32 and 64 for the experiments, which limited the performance accuracy. The system attained the highest speaker identification accuracy at 99.5% for the TIMIT database and 60.7% for the NTIMIT database. In addition, noise and handset effects were not included, which is a drawback of this paper. Similarly in 1995 [40], only TIMIT and NTIMIT databases were used to examine performance accuracy against the number of speakers for clean and telephone speech for the NTIMIT database. The identification accuracy obtained was 99.5% and 60.7% for the TIMIT and NTIMIT databases, respectively. In 1995, Reynolds and Rose [41], utilized the GMM for robust text independent speaker identification. The system was tested with conversational speech from the KING database, which included 51 male speakers. The evaluation of conversational telephone speech for 49 male speakers examined different issues such as variance limitation, initialization, and selection order in the model, and the GMM was compared with various speaker

### 2.3 Literature Review for Speaker Identification

---

modelling methods. The main drawback of this paper was the limited population size and that the effect of noise and the handset were missing.

In 2012, Kumari et al. [1] produced fusion based between MFCC features and their inverse features (IMFCC) for text independent speaker identification by employing the GMM to model speakers. The identification system was conducted with 120 speakers from Dialect Region (DR) one and four from the TIMIT database. The system achieved the highest identification rate of 93.88% at 16 Gaussian Mixture Component (GMC) size using weighted sum fusion. The drawbacks of this research were using a limited number of GMCs and only testing the system with clean speech; real environments such as noise conditions and handset effects were missing, and the study does not mention how many speakers were taken from DRs 1 and 4, respectively. In the current thesis, a higher identification rate was achieved compared with this paper, and different background noise were included with different feature and fusion methods, to improve the performance accuracy.

For GMM-UBM modelling, in 2011, Togneri and Pullella [3] presented an overview of two main issues: robustness and accuracy of speaker identification. The authors investigated the GMM-UBM system with 39 feature dimensions of the concatenated MFCC, delta, and acceleration vectors, by employing the CMN to remove the channel effects. The evaluation tested 64 speakers (32 male and 32 female) selected to balance gender and eight different dialect regions (for each dialect region, there were four male and four female speakers). The G.712 handset type was used in both training and testing phases and the system was modelled by 128 mixture components of GMM. The identification accuracy attained was 94.5% for clean speech with handset with presence of CMN, while the system degrade at white noise with 74.2% at 30 dB . The paper did not include different background noises, and also new technologies such as fusion, which could have improved the performance accuracy, were missing. In addition, evaluation with challenging and realistic noise was missing and there was also a limited population size.

In 2009, Apsingekar and Leon [42] exploited a multi-class SVM for speaker identification and compared it with the GMM-UBM approach as a baseline system, by evaluating with the NIST 2002 database. The highest number of speakers was 64 and the highest identification rate obtained was 97% based on SVM, compared with 98.44% from GMM-UBM. The paper had a limited number of speakers, and noise conditions were not considered.

### 2.3 Literature Review for Speaker Identification

---

Various other approaches have also been used to model speakers. In 2009, Revathi et al. [43] presented a clustering method for both speech and speaker recognition, and the highest identification rate was 91%. For speech recognition, TI *digits*<sub>1</sub> and TI *digits*<sub>2</sub> databases were exploited with isolated digits and continuous speech, respectively. However, 50 speakers from the TIMIT database were tested for speaker recognition. In this study, there was again a lack of speakers and realistic noise conditions.

In 2013, Bhardwaj et al. [44] exploited the Generalized Fuzzy Model (GFM) which used three forms for modelling speakers for speaker identification tasks: HMM-GFM, GMM-GFM, and fusion based HMM-GFM, in which HMM is the Hidden Markov Model. The system was tested with 40 speakers from the VoxForge speech corpus and 140 males from the 2003 NIST 2003 (NIST 2003). In addition, databases were studied with different SNRs (-5, 5, 10 and 20) dB by applying various noises such as babble, a car, a destroyer engine, and factory noise from the NOISEX database. The results for NIST 2003 were worse compared with the VoxForge database. The highest identification accuracy was achieved using HMM-GFM (fusion) with, respectively, 93%, 92%, 91%, 92% and 92% for clean, car, babble, destroyer engine and factory noise using the VoxForge database, compared with 51% for both GMM-GFM and HMM-GFM (fusion) on NIST 2003 database. Even though good results were achieved with the VoxForge database, there was still a lack of speakers, and indeed there are limited researchers who include the I-vector for modelling speakers, as discussed below.

In 2013, McLaren et al. [45] developed robust features for highly degraded speech via transmission channels for speaker identification, and then at a later stage these features were combined in the I-vector framework. In addition, the Hidden Markov Model (HMM) and GMM were utilized with the Speech Activity Detector (SAD) to extract the I-vector, and then Probabilistic Linear Discriminant Analysis (PLDA) was used to recognize speakers. To improve the system performance, both I-vector and score fusion were considered. The evaluation was interested in multiple durations for tests and enrolment (3, 10, 30, 120) seconds. The performance measure used the Equal Error Rate (EER) and the best EER at the evaluation condition of 10-10 seconds was 9.4%, carried out with both score and I-vector fusion. The drawback of this paper was the complexity of the system,

### 2.3 Literature Review for Speaker Identification

---

which used the EER instead of the identification accuracy. However, realistic background effects for different SNR levels were missing.

In 2014, Liu et al. [46] employed the I-vector with 400 dimension and three channel compensation methods for text independent speaker identification. The session compensations were: LDA, WCCN and Nuisance Attribute Projection (NAP). However, for classification purpose, the SVM and CDS were applied. The databases were a corpus designed by members of the author's laboratory and a voice library of MIT mobile phone speaker recognition. The sampling rate for the speech utterances was 8 KHz. The evaluation involved 50 speakers, 30 males and 20 females, using the same microphone. In addition, the length of the training data for each speaker was three minutes, and the testing data was 30 seconds. The highest accuracy rate was achieved using the I-vector + LDA + WCCN approach, with the CDS classifier at 94.14% . However, the authors presented two different comparisons in terms of the accuracy rate for different compensation methods and two classifiers (CDS and SVM), and there was again a lack of speakers. In addition, different realistic noise and environments conditions were missing. However, in the current thesis, a higher accuracy rate is achieved and includes various challenging environments with a larger number of speakers.

In 2014, Schmidt et al. [47] presented a fast retrieval approach which combined the Locality Sensitive Hashing (LSH) method with the I-vector approach through the cosine distance for speaker identification. LSH is a compromise between running time and accuracy. About a thousand single speakers from YouTube were studied, with at least half an hour of talk from each video. The best results in terms of relative accuracy are 92%, 96.1% and 98.4% for utterance lengths of 10, 20, 60 seconds, respectively. However, a standard database and challenging environments, including studies of different background noise, were not considered in this paper. Ultimately, the performance accuracy for 10s testing was lower than that achieved in this thesis at 8s testing, where all the above points were considered.

In 2014, Karadaghi et al. [48] investigated three different techniques: GMM-UBM; GMM-UBM with TZ-score normalization method; and I-vector with 300 dimension for text independent open set speaker identification. The database

### 2.3 Literature Review for Speaker Identification

---

employed was the NIST speaker recognition evaluation SRE 2008. A sub-set of the database of telephone-quality speech from 400 speakers and 200 unknown speakers (out-of-set) were selected, from short 2 and short 3 as core condition. For UBM training, 1,554 utterances from 932 females and 622 males (a subset from NIST SRE 2005 database) were used. In addition, white noise, factory noise, car noise from NOISEX-92 corpus were added to the telephone speech utterances of the NIST 2008 database, with three levels of Signal to Noise Ratios (SNRs) (5-15) dB. The first stage of the evaluation used 400 speakers for closed set speaker identification, while the second was for verification. The highest identification rate was 49.5% with the I-vector, which outperformed the GMM-UBM with and without score normalization for clean speech. Also, the paper showed that the I-vector approach is more robust than the GMM-UBM approach. The main drawback for this paper concerns the lack of information about the early system stages related to feature extraction methods, and feature dimensions. In addition, very poor accuracy was achieved even for clean speech, and more than 50% of the testing samples failed to be identified. This thesis used the microphone channel of the NIST 2008 database, which has not been considered, as well as four different databases, and realistic challenging noise databases, by providing different fusion techniques to improve the performance accuracy.

In 2016, Matjka et al. [49] analysed the I-vector based approach for speaker identification and employing the Bottleneck (BN) features of a Deep Neural Network, as well as the conventional MFCC features and their concatenation. The evaluation system tested the telephone condition of the NIST SRE 2010, the EER and the Detection Cost Function (DCF) were used for the performance measurements. The system did not show how far different background noises affected the performance accuracy. In addition, the paper did not show the effect of evolution on different databases with the same system.

In comparison to all the above research on speaker identification based on the I-vector approach, this thesis achieves a through evaluation in terms of UBM mixture sizes and SNR levels using four different databases. One of these databases is the 2016 challenge database with different fusion technologies. In addition, the system was evaluated for different challenging environments using the handset and various background noises. However, these were not considered in any

of the above studies.

### 2.3.4 Noise Robustness and Challenging Environments

In 1995, Reynolds et al. [50] presented a text independent speaker identification system which examined the effect of increasing the number of speakers and the degradations produced from telephone transmission. The system was implemented by changing the number of speakers to 630 for clean, wideband and telephone speech. The system was modelled by GMM and conducted with 630 speakers from the TIMIT and NTIMIT databases, respectively. SIA of 60.7% and 99.5% was obtained from NTIMIT and TIMIT databases, respectively. The paper also measured the performance loss by telephone transmission with a large population. This was achieved by regularly degrading the TIMIT speech to match with measured NTIMIT degradation, and then gauging the performance loss at each step. In addition, measuring the distortion of the non-linear microphone was also considered in this paper. Different background noise effects, such as AWGN and non-stationary noise, were missing. In addition, a database including realistic noise rather than just noise added was not considered .

In 1996, Reynolds [51] presented a study using the Switchboard corpus in experimental work on the effects of handset variability for text independent speaker recognition. To include the handset, the caller's telephone number was utilized for each conversation. The same telephone handset was expected to have been used in conversations produced by identical telephone numbers, while different handsets were assumed to have been used for different telephone numbers. Between the testing and training utterances, the first part of this study focused on the mismatch and match handset conditions using the SPIDER database. The second part employed the May95 NIST SRE database in terms of the handset variability. In addition, some channel compensation methods were applied. The empirical experiments included 160 imposter speakers (78 female, 82 male) and 45 claimant speakers (18 female, 27 male). Furthermore, in May 1995, the NIST database SRE derived 80 imposter speakers (47 female, 33 male) and 26 claimant (11 female, 15 male) from the Switchboard. The handset variability with one handset type was used for training and testing by different type of handsets. The study was limited to the handset effect, and other environments such as noise

### 2.3 Literature Review for Speaker Identification

---

conditions were not considered.

In 2007, Ming et al. [52] presented speaker identification and verification under noise conditions. The paper handled the worst-case scenario, which considered real-world applications such as handheld devices or the Internet, in terms of the fact that prior knowledge of the noise is not available. The authors provided a new training method to model the noise by combining a multicondition training model and missing information theory. In order to reduce the mismatch between training and testing, both coarse and smooth compensations were used. Coarse compensation for noise can be achieved by restricted noise variation, while smooth compensation was achieved using missing information theory, conducted outside the given training conditions by neglecting noise variation. However, in this paper to attain the optimum recognition performance with less model complexity, various training data were used. The speaker recognition system was conducted using re-recordings from the TIMIT database in the presence of different noise types, and a collected handheld device database in realistic noisy environments. The paper showed that the new model achieved better performance compared with the baseline systems. The drawback of this work was the lack of handset variability for the TIMIT database. In addition, there is a limitation with the identification accuracy, since only the Gaussian Mixture Components (GMCs) with  $\{32, 64, 128\}$  were used and increasing the GMCs might decrease the accuracy. In addition, AWGN was not considered in this paper, nor was a realistic database in which the noise had not been added artificially.

In 2009, Khanteymoori et al. [53] characterized both the implementation and theory of Dynamic Bayesian Networks (DBN). MFCC and Delta MFCC were used and the study employed DBN learning and compared it with the traditional GMM. Different background noises were applied to see the effect on speaker identification accuracy in presence of noise such as white noise, babble, and F16, and a factory, and the noise was taken from the NOISEX database with various Signal to Noise Ratios (5-20) dB with 5 dB step size. 50 out of 64 speakers were selected with respect to gender, age, dialect region, and educational level from a Farsi telephony speech database. Two speech datasets were employed; spontaneous data of three seconds was utilized for testing, as well as for training 30-second utterances from reading speech. The drawback was the limited population numbers and the omission of the handset effect.

### 2.3 Literature Review for Speaker Identification

---

In 2010, Wang et al. [22] presented an integration of the MFCC with phase information for speaker recognition. In clean speech, the error rate for the speaker identification was reduced to 78% using the phase information. Effectively, the phase information was efficient in noisy environments. However, in the presence of noise conditions, it was combined with the MFCC to decrease the identification error rate by 20% to 70% compared to MFCC alone. The system was tested with 35 speakers from the NTT database and 270 speakers from Japanese Newspaper Article Sentences. Many aspects of this study are of note, and one of these is the use of Japanese speakers, which means that no fair comparison with other work using standard language (English) can be made [22]. In addition, the paper does not include a realistic database such as NIST, and challenging environments, such as the handset effect, were missing.

In 2011, Togneri and Pullella [3] considered the effect of the G.712 type handset on closed-set speaker identification. The authors utilized 64 speakers, balanced for gender and dialect, using the GMM-UBM approach. The system achieved identification accuracy of 94.5% with Cepstral Mean Normalization (CMN) and the G.712 type handset. In addition, the system degraded when tested with AWGN and attained 74.2% at 30 dB in presence of both CMN and the handset effect. Other realistic background noise was not considered in this paper, and there was also a limited population size and an absence of new technologies, such as fusion.

In 2011, Wang et al. [54] used various techniques for denoising the effect of additive noise from MFCC features (vocal tract) and the Wavelet Octave Coefficients Of Residues (WOCOR), which represent the vocal source features. To remove the residual signal and then improve the robustness of the WOCOR, the frequency domain approach was employed. MFCC was calculated from enhanced speech by applying spectral subtraction to the MFCC features. The paper showed that using combined denoising for both WOCOR and MFCC were efficient in the presence of additive noise for speaker recondition. The performance measures were EER and the identification error rate. 50 male speakers from CU2C Cantonese speech database in Hong Kong at the Chinese University were tested, and the NOISEX-92 database was used to add noise. The paper focused on error rather than accuracy, but different realistic noises and the handset effect were missing, and there was also a lack of speakers.



### 2.3 Literature Review for Speaker Identification

---

In 2011, Li and Huang [55] presented a new auditory feature extraction method called Cochlear Filter Cepstral Coefficients (CFCCs). The main aim of this research was to apply the new features CFCCs for speaker identification to handle the mismatch problem effect between training and testing conditions (training using clean speech and testing by noisy speech). This paper illustrates the effectiveness of the CFCCs compared with MFCC in the presence of noises such as babble, a car, and white noise. Both features achieved the same identification accuracy for clean speech at 96%. However, at 6 dB CFCCs 88.3% was achieved, compared to 41.2% from MFCC. CFCCs therefore outperformed PLP and RASTA-PLP features under white noise, and had similar performance for babble and car noise compared with RASTA PLP. The evaluation for this paper was achieved by employing 460 speakers from the NTIMIT database for testing purpose and 38 speakers for the development phase, to show that CFCCs were better than MFCCs. In addition, 34 speakers from the Speech Separation Challenge (SSC) were utilized under mismatch conditions, and the data including several conditions. A realistic database for speaker identification, such as the NIST, was missing and can be used as alternative to the SSC database. Also, using SNRs limited to (0, 6) dB is not enough for studying the noise effect. New technologies such as fusion were also missing and could have exploited these features to improve the accuracy.

In 2013, Zhao and Wang [56] presented Gammatone Frequency Cepstral Coefficients (GFCCs) as a new and more robust feature compared with MFCC. This paper mainly focused on analysing robustness to noise for both GFCCs and MFCC features in terms of speaker identification. The paper also showed how to improve the robustness for both features. The evaluation was achieved by 330 speakers randomly selected from the TIMIT database. In addition, factory noise from the NOISEX-92 database was considered. The paper illustrates that the cubic root rectification produced by GFCCs is more robust compared with non-linear rectification represented by the log function in MFCC. The paper emphasized the comparisons in terms of the noise robustness between GFCCs and MFCCs features, rather than focusing on identification accuracy. However, testing with a challenging database as an alternative to the ideal TIMIT database was missing. In addition, studying noise effects such as the AWGN and different non-stationary types in presence of a handset were not included. However, the

### 2.3 Literature Review for Speaker Identification

---

fusion of both features might have improved both the robustness and the accuracy, which was not considered.

In 2014, Maged et al. [57] presented MFCCs features extracted from noisy speech with AWGN and Discrete Wavelet Transform (DWT) for robust speaker identification. To reduce the data, the Vector Quantization Linde-Buzo-Gray method was used. From degraded speech, feature extraction was employed by DWT to achieve higher identification accuracy, because the features in the approximation part of the DWT were added. The system was tested with eight and thirteen speakers under AWGN. The main drawback of this paper was the lack of speakers and the failure to use standard and realistic databases. This paper also did not cover handset and non-stationary noise.

In 2014, Zhao et al. [58] addressed reverberation and additive noise for robust speaker identification. In this paper, a deep neural network was employed to remove the noise over binary masking. Telephone conversation from the NIST 2008 was used to evaluate the system by utilizing 300 random speakers, as well as 50 speakers from the TIMIT database. However, this paper also did not cover a handset and, the newest state of the art I-vector approach could have been exploited instead of the GMM-UBM.

In 2016, Islam et al. [59] presented 2-D neurograms structured from the auditory periphery for a new speaker identification system. To construct the neurograms, speech signals were simulated for a broad range of characteristic frequency of responses in auditory-nerve fibres. The GMM-UBM algorithm was used to train the coefficients for the neurograms and model the speakers. The system was tested with the TIMIT, TIDIGIT, and YOHO databases for text independent speaker identification with 100, 40 and 137 speakers, respectively. In addition, 39 Malaysian native speakers from University Malaya (UM) database were enlisted for text dependent speaker identification. The system was tested with street, pink and white Gaussian noise for various SNR levels. Furthermore, MFCC, frequency domain linear prediction and Gammatone frequency cepstral coefficients were employed. However, the handset problem and a realistic database such as the NIST were not considered in this work. Moreover, the I-vector approach might have been proposed to improve the identification accuracy. New techniques such as fusion could also have been proposed to improve the identification accuracy. New techniques such as fusion could also have been

suggested to improve the identification accuracy.

### 2.3.5 Classification

In 2009, Apsingekar and Leon [42] presented the multi-class SVM as a classifier for speaker identification. This was compared with the Maximum Likelihood (ML) for GMM-UBM speaker modelling approach. Although, using a multi-class SVM classifier gives better identification accuracy than the (ML), this study still had a limited population size. In addition, their paper failed to consider the effect of different noise and channel conditions, and their effects on the used classifier.

In 2013, Hu et al. [60] presented a fuzzy clustering speaker identification system based on a decision tree under AWGN conditions and a large number of speakers. The system compared the base lines of MFCC with GMM and MFCC+GMM+UBM. The idea behind using a decision tree for classification is to enable the system to succeed with a large population under AWGN. The system was tested with 3,805 speakers collected online from the websites of audiobooks. However, even though the authors used a large number of speakers, the database was not standard for speaker identification. In addition, the system was not tested for different realistic noise conditions such as non-stationary noise and other challenging conditions. It would also have been helpful if the authors had used the fusion based I-vector approach for speaker identification for big data.

In 2014, Nidhyananthan and Kumari [61] presented the ML for GMM and a single hidden layer feed forward neural network, which exploited the ELM as a classifier. The system classified 50 speakers and showed that the ELM is 20 times faster than the GMM in the testing phase. However, the GMM outperformed the ELM and obtained identification accuracy of 94%, compared with 79.25% achieved by the ELM algorithm. In addition, it was found that the ML is effective for classification when combining MMFCC and MFCC and attained the highest speaker identification accuracy of 97.5%. This thesis shows that the ELM is efficient, fast and less time consuming for speaker identification when using the I-vector approach, and this combination has not been used to date.

In 2015, Almaadeed et al. [30] developed multimodal neural networks using Probabilistic Neural Network (PNN), Radial Based Function Neural Network (RBF-NN) and General Regressive Neural Network (GRNN). The feature

## 2.3 Literature Review for Speaker Identification

---

extraction for the identification system was based on different wavelet features, and the fusion was a decision level scheme. Major voting between three neural network classifiers was used to identify speakers. On this basis, the system was tested with 34 speakers, 16 females and 18 males, from the GRID speech corpus. Cross validation with 10 fold was employed and the best accuracy was 97.5% using wavelets based on the Multimodal NN (MNN). Again, the drawback with this study was using a limited number of speakers and failing to consider more challenging environments, such as noise and channel effects. However, in the current thesis, all these points were handled.

In 2015, Nandya et al. [62] presented an Artificial Neural Network (ANN) classifier for fixed text (text dependent) speaker identification based on MFCC features. A Back Propagation Neural Network (BPNN) was exploited to identify the individual voice characteristics of 50 users. The system achieved 92% identification accuracy. But had the main drawback of a low population size and a lack of consideration of different noise conditions and channel effects.

### 2.3.6 Fusion Technologies

In 2012, Kumar et al. [1] presented the weighted sum fusion between MFCC and the inverse of MFCC (IMFCC) to improve the identification accuracy. The system was tested with 120 speakers from the TIMIT database and the highest identification efficiency was 93.88%. In the current thesis, alternative fusion methods were proposed such as maximum, mean, interleaving, concatenated, and cumulative fusion methods, to improve the identification accuracy. Chapters 5 and 6 describe all the fusion methods in more depth. In addition, in Chapter 4, feature and score fusion were used to give different prospectives on improving performance accuracy.

In 2012, Nidhyananthan et al. [63] developed a combination of vocal tract features such as Dynamic Mel Frequency Cepstral Coefficients (DMFCCs) and MFCC to improve speaker identification accuracy. However, score fusion was also considered for the same purpose. The features extracted the spectral characteristics and the dynamic behaviour such as formant, bandwidth formant frequency, and pitch frequency. The system was modelled using GMM. The paper tested 630 speakers from the TIMIT database and the maximum accuracy

### 2.3 Literature Review for Speaker Identification

---

achieved was 89.78% and 92.53% for DMFCC and MFCC, respectively; the weighted sum fusion attained 98.02%. The drawback for this paper was that no realistic evaluation of different noise conditions and the handset effect were considered. In the current thesis, other fusion methods are considered: maximum, mean, concatenated, interleaving, and cumulative fusion methods, in order to improve the identification accuracy.

In 2012, Nakagawa et al. [64] combined GMM, the MFCC and phase information for speaker identification and verification tasks, as well as score fusion based on weighted sum. Both tasks were conducted using the NTT database with 35 Japanese speakers (13 female and 22 male) and JNAS database, which is a large-scale and has 270 speakers (135 female and 135 male). The best speaker identification performance obtained was 98.8% and 97.7% for the combination of modified phase information with the MFCC and the MFCC, respectively. However, the best speaker verification EER was 0.45% and 0.72% for the combination and for MFCC, respectively. The drawback for this paper is again that background noise conditions and channel effects were not considered. In addition, using Japanese speakers for speaker identification does not provide a fair comparison when the results are compared with other work for standard English language.

In 2013, McLaren et al. [45] produced a multiple fusion system based on a fusion I-vector framework using concatenated fusion and score level fusion. In this work, the performance measure was the EER for the speaker identification system. The system was very complicated, and the current thesis gives a new prospectives on fusion technologies. In addition, concatenated and weighted sum fusion, which can also improve the system performance, it considers interleaving, cumulative, maximum, and fusion mean. Furthermore, McLaren et al. could also have include noise effects in the system, while are considered in this thesis.

In 2014, Nidhyananthan and Kumari [61] presented an Extreme Learning Machine and a Gaussian Mixture Model for text independent multi-lingual speaker identification. The system utilized MFCC, Modified MFCC, BFCC and Linear Predictive Residual Cepstral Coefficient (LPRCC). The system showed that the GMM outperforms ELM and achieved 97.5% in speaker identification accuracy, with a 40 filter bank size at frame size 256 and 128 frame shift. The system was tested with a jyamagis tool kit (synthesized voices) and recorded voices. 50 speakers were used, some of which were recorded and some of which was

synthesized speech selected from the jyamagis tool kit. Score level fusion was applied and attained the highest speaker identification accuracy at 97.5%, by combining the scores between MMFCC and MFCC by using the weighted sum fusion. However, the drawback of this paper was its failure to consider both channel and noise effect, and different fusion technologies such as those presented in this thesis.

## 2.4 Summary

This chapter was organized as following: Section 2.1 included the auditory perception system of the human communication system, and then the auditory nervous system and the cochlea. Section 2.2 gave an overview of the speaker identification system used in this thesis and focused on feature extraction methods and the normalization techniques. However, other stages for the system will be discussed in the contribution chapters. Section 2.3 described the literature review in terms of the speaker identification task. This section can be categorized into six subsections based on: feature extraction, I-vector extraction, modelling, noise robustness, and challenging environments, classifiers and fusion techniques. Section 2.4 summarised the chapter.

The literature review in this chapter covers 54 studies; 12 were on feature extraction, and six on I-vector extraction. Furthermore, 14 were on modelling speakers, and 5 of these were on the I-vector model; 9 out of the 14 were on other modelling approaches. Moreover, 12 studies were included on different challenging environments, such as background noise (AWGN, Non-Stationary Noise), handset, and reverberant. In addition, five studies covered classifiers and five fusion methods were present in more than one subsection. According to Fig. 2.8, the pie chart represents the distribution within the literature review of the different speaker identification stages in 54 studies, measured by their Percentage proportion. The distributions were 22.22%, 11.11%, 25.93%, 22.22%, 9.26% and 9.26% for feature extraction, I-vector extraction, modelling, environments, classifiers and fusion, respectively. In addition, in the modelling based section, 9.26% was on I-vector based speaker identification and 16.67% on other modelling approaches. Only nine percent of the 54 studies included the I-vector approach to speaker identification, and furthermore these studies did not include challenging

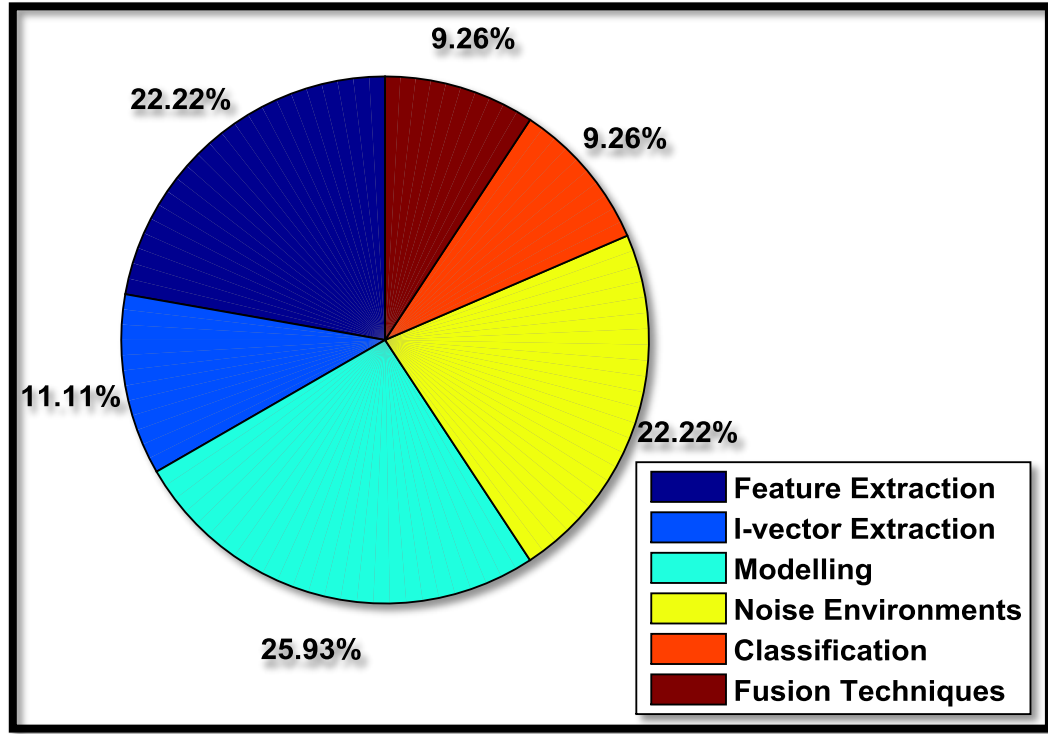


Figure 2.8: Percentage Proportion of Each Stage of the Speaker Identification System in 54 Studied in the Literature Review Based on Six Stages of the System: Feature Extraction, I-vector Extraction, Modelling, Noise Robustness, and Challenging Environments, Classifiers and Fusion Techniques.

environments, such as different stationary and non-stationary noise types or handset effects. In addition, studies have not considered different databases with and without fusion techniques, and the novel contribution of this thesis is that all these deficits are addressed. The next chapter will discuss various databases, some of which are used in this thesis, and the reasons for avoiding certain databases.

# Chapter 3

## Databases and Performance Measurement

### 3.1 Background

This chapter consists of two major parts: the first focuses on the databases, while the second considers performance measurement. In part one, due to their different characteristics, four main databases are used in this thesis: the TIMIT; the 2016 SITW Speaker Recognition Challenge; the 2008 NIST; and the NTIMIT, which is a telephone bandwidth version of TIMIT. In addition, various databases are discussed in terms of the reasons for excluding these databases, according to their drawbacks for this study. In part two, performance measurement is employed to evaluate the speaker identification systems, and the SIA is considered for all contribution chapters in this thesis. In addition, Detection Error Tradeoff (DET curve), EER and minimum DCF for speaker recognition are discussed in this chapter, especially for the verification application, although they were not employed in this thesis.

### 3.2 Databases

In this section, two major types of databases are discussed: type 1 includes the four databases used in this thesis, while type 2 covers other databases not considered. One of the most important databases is the TIMIT database, which has various types and versions that are widely used for different state of the art applications. Therefore, before discussing type 1 and 2 separately, the TIMIT



## 3.2 Databases

---

family of databases is discussed in this section, including some examples of both types 1 and 2 databases. TIMIT is a speech corpus comprised of male and female American English speakers with different dialects; in the database, the speech has been lexically and phonemically transcribed. There are different versions of the TIMIT database, which are called in this section the TIMIT Family:

- TIMIT Acoustic-Phonetic Continuous Speech Corpus-1993
- NTIMIT-1993
- CTIMIT-1996
- FFM-TIMIT-1996
- HTIMIT-1998
- MOCHA-TIMIT-1999
- The VidTIMIT Audio-Video Dataset-2001
- STC-TIMIT 1.0-2008
- WTIMIT 1.0-2010
- TCDTIMIT-2015
- Noisy TIMIT Speech-2017

A brief description of the TIMIT family databases is summarized in Tables 3.1, 3.2 and 3.3 below:

The tables are categorized into the two columns: the TIMIT type and the description for each type. The description for eleven types of TIMIT databases include: the data source, sample type, sample rate, applications, aims, authors, number of speakers, language and further information from the websites.

Table 3.1: Description of the TIMIT Family of Databases-Part A

Type of TIMIT	Description
1- TIMIT Acoustic-Phonetic [65] Continuous Speech Corpus	Data source: microphone speech Application: speech recognition Sample Type: 1-channel pcm Sample Rate: 16 kHz, Language: English Authors: John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, Victor Zue Number of Speakers: 630 of 8 major dialects of American English, Member Year: 1993
2- NTIMIT is [66] (Network TIMIT) a telephone bandwidth of TIMIT corpus	Data source: telephone speech, Application: speech recognition Sample Type: 1-channel pcm Sample Rate: 16 kHz Authors: William M. Fisher, George R. Doddington, Kathleen M. GoudieMarshall, Charles Jankowski, Ashok Kalyanswamy, Sara Basson, Judith Spitz Number of Speakers: 630 with 6,300 original TIMIT recordings through a telephone handset and over various channels in the NYNEX telephone network Member Year: 1993
3- CTIMIT is Cellular bandwidth [67] to the TIMIT Speech Corpus	Data source: telephone speech, Application: speech recognition Sample Type: 1-channel pcm Sample Rate: 8 kHz, Language: English Authors: E. Bryan George, Kathy L. Brown, Martha Birnbaum, Michael Macon Aim: to provide a large database to exploit for the evaluation and design of the operating systems for speech processing in varied cellular telephone environments American English, Member Year: 1996
4- FFMTIMIT Far Field [68] Microphone Recording Version	Data source: microphone speech, Applications: speech recognition Sample Type: 1-channel pcm, Sample Rate: 16 kHz, Author: John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, Victor Zue FFMTIMIT contains the secondary microphone waveforms for TIMIT database. The primary were recorded using a close-talking noise-cancelling head-mounted Sennheiser microphone (model HMD414). The secondary was a Breul and Kjaer (B and K) 1/2" free field microphone (model 4165). Member Year: 1996

Table 3.2: Description of the TIMIT Family of Databases-Part B

Type of TIMIT	Description
5- HTIMIT [69] Handset TIMIT, which contains a subset of 192 female and 192 male speakers through different telephone handsets	Data source: telephone speech, Applications: speech recognition, speaker identification Sample Type: 1-channel pcm Sample Rate: 8 kHz Author: Douglas Reynolds Aim: to study handset transducer effects on speech recognition systems, Ten transducers (telephone handsets) were used, Member Year: 1998
6- MOCHA-TIMIT [70]	Data source: microphone speech, Sample Rate: 16 kHz Authors: Alan Wrench, Queen Margaret University College When created: November 1999, Corpus: a set of 460 short sentences for British-TIMIT sentences
7- VidTIMIT Audio-Video Dataset [71]  is comprised of video and corresponding audio recordings of short sentences from 43 people	Data source: Audio-Video For useful topics: automatic lip reading, multi-modal speech recognition, multi-view face recognition and person identification Number of people: 43 people Author: Conrad Sanderson The dataset was recorded in 3 sessions. The sentences were selected from the test phase of the TIMIT/NTIMIT corpus. Each person has ten sentences. Session 1 includes the first six sentences. Session 2 includes 2 sentences and also Session 3. Each session from each person has a head rotation sequence: Moving their head to the right, left, back to the center, down, then up, and back to the center. The recording was made in an office environment. A digital video camera of broadcast quality was used, at a resolution for each person of 512 x 384 pixels. JPEG images have a setting quality of 90%. Audio files are stored as 16 bit, mono, 32 kHz. Years: 2001-2002, Website available: <a href="http://conradsanderson.id.au/vidtimit/">http://conradsanderson.id.au/vidtimit/</a>
8- STC-TIMIT 1.0 [72] is a telephone version of the TIMIT database	Data source: telephone conversations Application: speech recognition, speech synthesis Sample Type: ulaw, Sample Rate: 8 kHz The training partition has 4,620 files, and the test partition has 1,680 files Two calibration tones with four sets were generated 2 sec. 1kHz tone, 2 sec. sweep tone from 10 Hz to 4000 Hz. Authors: Nicolas Morales Member Year: 2008

Table 3.3: Description of the TIMIT Family of Databases -Part C

Type of TIMIT	Description
9- WTIMIT 1.0 wideband mobile [73] telephony derivative from TIMIT database	Data source: telephone speech, Application: speech recognition, speaker identification Sample Type: 1-channel signed linear PCM Sample Rate: 16 kHz Authors: Patrick Bauer, Tim Fingscheidt The testing subset contains 1,680 speech files. The training subset consists of 4,620 speech files. Member Year: 2010
10- TCD-TIMIT [74]: An audio-visual corpus of continuous speech	Data source: audio-visual Consists of high-quality video footage and audio Number of speakers: 62 speakers In total, there are 6,913 phonetic sentences. The database is freely available for research use Application: audio-visual speech recognition research to help develop new approaches for the state of the art. To test the hypothesis using lip-speakers; three of the speakers are trained. In video footage for two angles were recorded with 30 and degrees straight on Aim: To create a new continuous audio-visual corpus designed for speech recognition research Years: 2015
11- Noisy TIMIT Speech [75]	Data source: microphone speech Application: speech recognition Sample Type: flac Sample Rate: 16 kHz Authors: Azhar Abdulaziz, Veton Kepuska The additive noise is: white, pink, blue, red, violet and babble noise with noise levels varying in 5 dB (decibel) steps and ranges from 5 to 50 dB Noisy TIMIT Speech was developed by the Florida Institute of Technology and contains approximately 322 hours of speech from the TIMIT database and modified with different additive noise levels Member Year: 2017

The tables show that the TIMIT and NTIMIT databases are the best databases in terms of the number of speakers, application and the availability, and hence the TIMIT and NTIMIT were selected for the work in this thesis.

### 3.2.1 Type 1: Databases Used in This Thesis

Four types of database were used in this thesis: the TIMIT, NTIMIT, SITW and NIST 2008 databases.

#### 3.2.1.1 TIMIT Acoustic-Phonetic Continuous Speech Corpus-1993

Although the TIMIT corpus is common and widely used [3], and essentially designed for automatic speech recognition systems, in this thesis it is exploited for identification purposes. The corpus' name is Texas Instruments (TI) and Massachusetts Institute of Technology (MIT). The TIMIT corpus contains 630 speakers, and each speaker has ten sentences, making a total of 6,300 sentences; the speakers were selected from eight dialect regions in the United States [65]. In this thesis, 120 of the 630 speakers were selected from two Dialect Regions (DR) namely, DR 1 and DR 4. 49 of the speakers were taken from DR 1 and the remaining 71 from DR 4, to match the work in [1] and [76]. Ten speech utterances were employed for each speaker; in the training phase, six utterances were used, and the remainder were kept for testing. A fixed speech length of 8 seconds (129,250 samples) was developed for all 120 speakers (1,200 speech utterances), and concatenation was used when necessary. In addition, further details about the TIMIT database concerning speaker distribution, text material and training/test partitions can be found on the website [77] and also the documentation for TIMIT on [65]. Firstly, Table 3.4 shows the TIMIT corpus speaker distribution, and the number of speakers according to sex given in parentheses. In the whole database, speakers who were children came from eight dialect regions represented in the geographical areas for the U.S, but in this thesis only DR 1 and DR 4 were exploited for speaker identification. Secondary, according to Table 3.5, there are ten sentences for each speaker: two sentences were from the Dialect (SA), then five from the Compact (SX), and the remaining three sentences from the Diverse (SI). The SA sentences were intended to disclose the dialectal variants of the speakers, while the SI sentences were chosen from existing text sources. To provide good

## 3.2 Databases

Table 3.4: TIMIT Dialect Region Distribution of Speakers

U.S. area for each DR	Dialect Region (DR)	Male	Female	Total
New England	DR 1	31 (63%)	18 (27%)	49 (8%)
Northern	DR 2	71 (70%)	31 (30%)	102 (16%)
North Midland	DR 3	79 (67%)	23 (23%)	102 (16%)
South Midland	DR 4	69 (69%)	31 (31%)	100 (16%)
Southern	DR 5	62 (63%)	36 (37%)	98 (16%)
New York City	DR 6	30 (65%)	16 (35%)	46 (7%)
Western	DR 7	74 (74%)	26 (26%)	100 (16%)
Army Brat	DR 8	22 (67%)	11 (33%)	33 (5%)
Overall	8	438 (70%)	192 (30%)	630 (100%)

Table 3.5: Speech Material for TIMIT Corpus

Sentence Type	Sentences	Speakers	Total	Sentences/Speaker
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total	2342		6300	10

coverage of pairs of phones, the SX sentences were designed. Finally, in this thesis' training and testing phases, partitioning was employed for six speech utterances for training (two sentences from SA + three sentences SI sentences + one SX sentence), and the rest (four SX sentences) for the testing phase to act as a mirror to the work in [1] in terms of the number of training and testing samples for each speaker.

Table 3.6: The NTIMIT Database Files Lacking Speech Data

path/name	nature of problem
ntimit/test/dr1/fdac1/sx394.flac	(0's in last 80%)
ntimit/test/dr1/fjem0/si634.flac	(0's in last 40%)
ntimit/test/dr2/mccs0/si839.flac	(mostly 0; first 13 csec is noise)
ntimit/test/dr5/fmah0/sx29.flac	(0's in last 60%)
ntimit/test/dr5/fmah0/sx299.flac	(0's throughout)
ntimit/train/dr4/mjdc0/si1161.flac	(0's throughout)
ntimit/train/dr7/fkde0/sx331.flac	(0's in last 60%)
ntimit/train/dr7/mjdg0/si1042.flac	(0's in last 90%)

### 3.2.1.2 NTIMIT-Network TIMIT

This database was produced by the Linguistic Data Consortium and developed by NYNEX Science and Technology Speech Communication Group [66] and [78]. This corpus presents a telephone bandwidth to assist the common TIMIT Acoustic-Phonetic Continuous Speech Corpus, and the data were contributed by NYNEX to NIST for distribution. 630 speakers with 6,300 speech utterances were transmitted over different channels by the NYNEX telephone network, and collected by NTIMIT. In this thesis, the same setup setting selected by TIMIT was used as well in the NTIMIT database in terms of: sampling frequency, dialect regions, number and name of speakers and speech utterances, and training and testing partitioning. After 1993, the NTIMIT database reported problems in the data file content, but no further information or corrections have been recorded by [78]. There are eight incomplete speech files, as explained in the Table 3.6. According to the Table 3.6, two speech files for DR 1 are problematic for both training and testing phases for two speakers from DR 1, and thereby could possibly reduce the recognition rate.

Table 3.7: NTIMIT Database Files Lacking Calibration Data

<b>path/name</b>	<b>BIN</b>	<b>LTU</b>
ntimit/test/dr3/fkms0/sx50.wav	01	0006
ntimit/test/dr2/fjas0/sx140.wav	01	0072
ntimit/test/dr5/mlih0/sx373.wav	01	0072
ntimit/test/dr7/fcau0/si1037.wav	01	0072
ntimit/test/dr7/mnjm0/si950.wav	01	0072
ntimit/train/dr1/mpsw0/sx437.wav	01	0072
ntimit/train/dr2/mbjv0/sx257.wav	01	0072
ntimit/train/dr3/mdhs0/si1530.wav	01	0072
ntimit/train/dr5/ftlg0/sx33.wav	01	0072
ntimit/train/dr7/fvkb0/si529.wav	01	0072
ntimit/train/dr7/mcre0/si1725.wav	01	0072
ntimit/test/dr2/fjre0/si1746.wav	01	0095
ntimit/test/dr2/mtas1/sx118.wav	01	0095
ntimit/train/dr1/fvfb0/sx222.wav	01	0095
ntimit/train/dr1/mrws0/sa1.wav	01	0095
ntimit/train/dr1/mwad0/sx72.wav	01	0095
ntimit/train/dr2/faem0/sx222.wav	01	0095
ntimit/train/dr2/fmjb0/si1177.wav	01	0095
ntimit/train/dr4/marw0/si646.wav	01	0095
ntimit/train/dr4/mbma0/si1222.wav	01	0095
ntimit/train/dr4/mlbc0/sx339.wav	01	0095
ntimit/train/dr7/mhbs0/si2205.wav	01	0095
ntimit/train/dr7/mmdg0/si1780.wav	01	0095

In addition, another serious problem was identified with 23 utterances, in terms of not knowing the precise channel effects for files identified with uncalibrated circuits. These 23 speech files are listed in Table 3.7, according to circuit ID for each file (“LTU”). From Table 3.7, it is evident that from the seven speech files, four belong to the DR 1, while the remaining three files are from DR 4. The problems in these utterances might be effect on the identifying their speakers and hence on the SIA.



### 3.2.1.3 The Speakers In The Wild Speaker Recognition Challenge 2016

This challenging database was fairly recently collected, and its main aim is to develop novel state of the art algorithms to benchmark existing technologies for researchers working in speaker recognition under various environments produced in the SITW database. Furthermore, this database is free and publicly available for research to those working in speaker recognition as explained in [79], [80]. The database was exploited and described using the following references [81], and [82], and then in addition to speaker recognition applications, various diarization algorithms were employed to exploit the new database [83]. The database is open source media, and has hand annotated speech samples from the database, including single and multi-speakers audio obtained in wild conditions. The SITW database contains about 300 persons under various conditions, such as: outdoors, in a stadium, red carpet interviews, clean interview conditions, and multi-speaker scenarios. In addition, the speech for each person was obtained using mobile phones and camcorders and is void of professional editing. In the SITW database, hundreds of individuals performed in a range of challenging environments, whereas in the video, speaker identities were visually confirmed. In this corpus, all compression, reverb, noise and artifacts were natural characteristics of authentic audio. However, with this real-world data and the many different challenging, it is expected to be difficult to recognize speakers in such conditions. In the thesis, 120 speakers were chosen; most of the selected speakers were single speakers and several were unbalanced multi-speakers. To select the target single speaker, Goldwave and Audacity software were employed to represent the single speakers considered in this thesis. In addition, to mirror the work in [1], each recorded speech was divided into ten equal lengths, of 8 seconds (129,250 samples) fixed length. However, to achieve the same fixed length of 8 seconds for all speech utterances, some were concatenated, and then six files were utilized for training and four for testing.

### 3.2.1.4 2008 NIST Speaker Recognition Evaluation Training Set Part 2-2011

This database was developed by NIST and Linguistic Data Consortium (LDC), and the authors were the NIST multi-modal information group in 2011, available

## 3.2 Databases

---

on [84]. This database handles telephone speech, with approximately 523 hours, and microphone speech, of about 427 hours, for both multilingual speakers using: English, Italian, Spanish, Arabic, Egyptian Arabic, Moroccan Arabic, Russian, Georgian, Uzbek, Iranian Persian, Persian, Urdu, Hindi, Panjabi, Tigrinya, Thai, Tagalog, Dari, Korean, Central Khmer, Northern Khmer, Bengali, Vietnamese, Yue Chinese, Wu Chinese, Min Nan Chinese, Lao, Mandarin Chinese, Chinese, Japanese. The sample type was u-law and the sampling rate was 8 kHz. As such, this database is intended for those generally interested in text independent speaker recognition. This database includes 950 hours of English language interview and multilingual telephone speech, with transcripts for training data in the Speaker Recognition Evaluation (SRE) to the 2008 NIST. The segments of telephone speech are summed-channel recordings and have length five minutes which are shorter than the longer original conversations. However, each speech conversation has two sides, the target and non target speakers where both participate and summed together. The interview material consists of segments of single channel conversation of at least eight minutes in length, taken from a longer interview. In addition, silence intervals were not deleted. In this thesis, the data source used is a microphone speech of native and bilingual English speakers in an interview scenario. The sampling frequency was converted from the original 8 kHz to 16 kHz for each speech file, and 120 speakers of English on a microphone channel were selected for comparison with the SITW and the TIMIT databases. However, only the single speakers were selected and the interviewers were removed. A fixed speech duration of eight seconds was created for each speech utterance, and four speech recordings were used for the testing phase and six for the training phase. Additional documentation is obtainable from the 2008 SRE for NIST website, also within the evaluation plan to the 2008 SRE.

### 3.2.1.5 Non Stationary Noise Database

Non-Stationary Noise (NSN) was used on the testing side only, and is available from [85] and [86] and. Both NSN and AWGN were adapted to fit the speech utterances by trimming them to the same fixed length of eight seconds (129,250 speech samples). In this thesis, three background noise types (NSN types) and AWGN with varying SNRs were tested for: street traffic, the interior of a bus, and a crowd environment. However, corresponding to the noise power (0dB to 30dB), there were seven SNR levels with 5 dB step size for each level.

### 3.2.2 Type 2: Databases Not Used in This Thesis

First of all, from the TIMIT Family, only TIMIT and NTIMIT were employed in this thesis; the other TIMIT family members not used in this thesis are: CTIMIT, FFM-TIMIT, HTIMIT, MOCHA-TIMIT, the VidTIMIT Audio-Video Dataset, STC-TIMIT, WTIMIT, TCDTIMIT and Noisy TIMIT Speech. However, these databases were not chosen, either because of issues in terms of the number of speakers, or the type of files and their availability. As well as the TIMIT Family databases, there are other databases which were not used in this thesis, such as: MOBIO, the GRID audiovisual sentence corpus, VoxForge, YOHO and the NIST I-vector Machine Learning Challenge Databases, which include: the Speaker I-vector Machine Learning Challenge, and the Language I-vector Machine Learning Challenge. In addition, the MATLAB Audio Databases Toolbox is also available but was not used in this thesis.

#### 3.2.2.1 MOBIO Database

This database provides a bi-modal database with audio and video forms for 152 individuals, 100 males and 52 females, and the data are available in [87] and also used in [88]. It was collected within approximately two years (August 2008 - July 2010) across five countries, but this led to a diversity of English speakers in the database in terms of native and non-native. In addition, two mobile devices were used with laptop computer (standard 2008 MacBook) and mobile phone (NOKIA N93i mobile) to record this database. The database has 12 sessions in total for each client, six sessions for Phase II, and the rest for Phase I. Phase II data include 11 questions and the range for the question types is short response questions, set

## 3.2 Databases

---

speech, and free speech. However, the Phase I question types range from short response questions, short response free speech, set speech, and free speech. This phase includes 21 questions. The drawback for this database is that the data source comes from a mobile and laptop, which makes it difficult to obtain a fair comparison with other databases; therefore, this database was excluded from this thesis.

### 3.2.2.2 The GRID audiovisual sentence corpus

This database is a high-quality corpus of 34 talkers, 16 female and 18 male. These audiovisual data contain both video (facial) and audio recordings in a thousand sentence corpus. The main aim is to assist in speech perception in common behavioural and computational studies. The data are freely available for research purpose [89]. In addition, for each talker, all video, audio and word transcriptions, as well as other associated information, are separately available. The absolute maximum amplitude value was limited for all audio files of “One” and also down sampled to 25 kHz where these have end pointed; the original 50 kHz signals raw data were also considered. In addition, two video file formats are presented, of high and normal quality, with  $(720 \times 576; 6\frac{kbit}{s})$  and  $(360 \times 288; 1\frac{kbit}{s})$ , respectively. Moreover, it is reported that the video of speaker 21 was unavailable due to a technical oversight. The main drawback for this database is also the lack of the speakers (only 34 speakers), and therefore this database was not considered in this thesis.

### 3.2.2.3 VoxForge Database

VoxForge is a free speech database and Open Source Speech Recognition Engines (OSSRE) and is free available as explained in [90]. This database was established to collect transcribed speech, and hence a free GPL speech corpus was produced and utilized with the open source speech recognition engines (on Linux, Windows and Mac). In addition, the acoustic models were compiled using the speech audio files, and then used for OSSRE, for instance ISIP, Julius (github), and HTK and Sphinx (note: HTK has distribution restrictions). Furthermore, LibriVox as a source of audio data, the VoxForge has been used since 2007. The major drawback of this database in terms of this thesis was very short speech length, and so it was not employed here.

### 3.2.2.4 YOHO Database

This database is a high-quality, large scale speech corpus initially proposed for text-dependent speaker verification tasks, for example in secure access technology. The sample type is 1-channel pcm compressed, while the sampling rate is 8 kHz and the data source was English microphone speech. These data were developed by the LDC, 1994 [91]. Through a US Government contract, this database was collected in 1989. In addition, the corpus was collected over a three month period from 108 males and 30 females. Further information can be seen in [92], [93], [94] and [95]. This database is not freely available and thus was not considered in this thesis.

### 3.2.2.5 NIST I-vector Machine Learning Challenge Databases

This database includes different challenging areas related to the Speaker I-vector Machine Learning and the Language I-vector Machine Learning Challenge, as explained in [96], from which the NIST I-vector 2014 database was produced. The NIST Speaker Recognitions from 2004 to 2012 is an I-vectors database, and each vector has 600 components derived from conversational telephone speech data. However, the database is classified into four file types: one file is a table representing the target speaker models; the other three files are three tables of I-vector development, model, and test I-vectors. Each table includes several rows, and each row includes: the corresponding speech length in seconds in order to compute the I-vector, the i-vector ID, and the 600-dimensional i-vector. It seems this database has ambiguity about the feature type and all system details and this is the main drawback for this challenging database. In addition to this limitation, the I-vector is of only 600 dimension and this requires a huge number of speakers and training utterances. Therefore, this database was not considered in this thesis.

### 3.2.2.6 MATLAB Audio Databases Toolbox

The MATLAB for Audio Database Toolbox (ADT) gives simple accessing and filtering as an alternative for manual custom coding and filtering, which is always needed for accessing the various databases by their metadata for instance YOHO, TIMIT. This avoids the time-consuming need to learn the database structure, and thereby allows concentration on the arithmetical aspects. In addition, the following databases are supported [97]:

- TIMIT Acoustic Phonetic Continuous Speech Corpus (AmericanEnglish)

### 3.3 Performance Measurement

---

- NTIMIT Telephone Network Acoustic Phonetic Continuous Speech Corpus
- CTIMIT Cellular Telephone Acoustic Phonetic Continuous Speech Corpus
- YOHO Speaker Verification Corpus
- TIDigits SpeakerIndependent recognition of connected digit sequences
- Children Voices Hebrew Speech
- Hebrew BGU Hebrew word samples
- Gutenberg Books MP3 format books

The supported search criteria for the TIMIT and NTIMIT databases are: speaker and sentence, dialect, sex, usage, phoneme, and word. However, the YOHO database supported: speaker, usage, numbers, and session, while the TI-Digits supports: speaker, digit, type, group and usage.

## 3.3 Performance Measurement

Generally, in the speaker recognition which includes both the speaker identification and speaker verification tasks there are two parts to the performance measurements: Part A, the EER and Part B, the SIA.

### 3.3.1 Part A: DET Curve, EER and min DCF

This part is essentially used in speaker authentication (speaker verification). In recognition systems such as speaker, image, biometrics and pattern recognition, two major types of errors can be classified, namely False Acceptance (FA) and False Rejection (FR) (further details are discussed in [98] and [99]). False Acceptance occurs when an imposter is successfully verified in error, whereas False Rejection happens when the true user is rejected. However, the rates for FA and FR are computed by the following formulas, as explained in equations 3.1 and 3.2 (employed in [98] and [99]):

$$FAR = \frac{\text{Number of FA}}{\text{Number of impostors accesses}} \quad (3.1)$$

### 3.3 Performance Measurement

---

$$FRR = \frac{\text{Number of FR}}{\text{Number of target accesses}} \quad (3.2)$$

Both rates (False Acceptance Ratio-(FAR) and False Rejection Ratio-(FRR) ) are based on the decision threshold. There are fewer false acceptances, while false rejections are more common when the decision thresholds are higher. On the other hand, there is more FA compared to fewer false rejections for lower decision thresholds, and so to compromise between the two operating rates, a decision threshold is used. In addition, DCF is a fixed decision threshold used to measure speaker verification performance; this decision is used to optimize the cost. This can be achieved by a weighted sum of the FR and FA rates by what is called the DCF. These weights correspond to probability of the target speaker,  $P_{tar}$ , which is also equal to  $P_{FR}$   $P_{tar} = P_{FR}$  and the a priori probability of impostor,  $P_{imp}$ , which is equal to  $P_{nontarget}$ ,  $P_{imp} = P_{nontarget} = P_{FA}$  trials. The costs  $C_{FR}$  and  $C_{FA}$  associated with the FRR, FAR, respectively, and the detection cost function, is determined by the following formula [98] and [99]:

$$DCF = C_{FR} P_{tar} FRR + C_{FA} P_{imp} FAR = C_{Miss} P_{Miss} P_{FR} + C_{FA} P_{imp} P_{FA} \quad (3.3)$$

where:  $P_{tar} = 1 - P_{imp} = P_{Miss}$ ,  $P_{imp} = P_{NonTarget}$   $C_{FR} = C_{Miss}$ ,  $C_{FR} = C_{FalseAlarm}$ . Based on the value of decision threshold, the value of DCF is dependent. When the decision threshold is changed, the minimum value of the DCF is acquired (Min DCF). Furthermore, the EER is the point at which the false acceptance rate is equal to the false rejection rate ( $FAR = FRR$ ), represented in the DET curve at the operating point. Moreover it could be used as another criterion to compare the speaker verification systems regarding the performance of these systems [98] and [99]. Moreover, in the DET Curve, both EER and DCF give speaker verification performance that corresponds to the operating point for a fixed decision threshold. To view the performance at different points, another method is used for the same curve that is DET curve. Instead of the Receiver Operating Characteristics (ROC) curve, the DET curve plots the variation for both the FA and FR rates, corresponding to different decision thresholds.

### 3.3.2 Part B: Speaker Identification Accuracy

The SIA is used to measure the percentage accuracy of the number of genuine speakers identified (true speakers identified) compared with the total number of speakers, as explained in (3.4) [3], [41] and [42]. The performance accuracy can be calculated by the SIA, represented by (3.4); some authors have exploited the SIA in [1] [100]:

$$SIA = \frac{\text{Number of True Speakers Identified}}{\text{Total Number of Speakers}} \times 100\% \quad (3.4)$$

In this thesis, the SIA is considered for all experiments in the next contribution chapters.

## 3.4 Summary

This chapter was organized as follows: Section 1 included the background; Section 2 dealt with the databases, including both the databases used in this thesis and those which were excluded, and why; Section 3 consisted of the performance measurement; In this chapter, four main databases (TIMIT, NTIMIT, SITW and NIST 2008) were presented, as well as the performance measurements using the SIA. These were all clearly and deeply discussed for all the experiments included in the following contribution chapters.



## Chapter 4

# Speaker Identification Using GMM-UBM Approach With Fusion and Evaluated on Original Speech Recordings

In this chapter, a new combination of features and normalization methods is investigated for text independent speaker identification. MFCCs features are efficient for speaker identification in original speech recordings, while PNCCs features are robust for noisy environments. Therefore, combining both features together is expected to be better than taking each one individually. In addition, CMVN and FW are used in order to remove or mitigate possible linear channel effects, they are also robust for channel and handset mismatch and additive noise in voice measurements. Speaker modelling is based on a GMM with a UBM. Coupled parameter learning between the speaker models and UBM is utilized to improve performance. Four main simulations for SIA are presented in this chapter including different fusion strategies: Late fusion (score based), early fusion (feature based) and early-late fusion (combination of feature and score based), late fusion for concatenated static and dynamic features (features with temporal derivatives such as first order derivative delta and second order derivative delta-delta features which are called acceleration features) and finally the statistically independent normalized scores for all the previous scores. 120 speakers from the TIMIT database are used in order to evaluate closed set speaker identification and to

mirror with [1]. Each speaker has ten speech recordings; six of them are used for the training phase, while the remaining four are applied for testing purpose as in [76] and [101]. A fixed speech length with 129250 samples (8 seconds for both training and testing) is achieved for each utterance via concatenation [76].

## 4.1 Background

Speech is commonly exploited as a human biometric due to the unique characteristics of individual voices [2]. In speaker identification and verification tasks, choosing the best features to capture this information is one of the most important issues. MFCC features are widely used for this purpose [3] [4]. However, to improve the SIA, MFCC features were fused with inverse MFCC features (IMFCC) in [1], but the approach was limited by the number of GMM components and the improvement in the recognition rate is still low. In addition, [64] proposed combining phase information with MFCC features to improve speaker identification. According to [62], a text independent speaker identification system can be achieved by using MFCC features and by using Back-Propagation Neural Networks (BPNNs) for classification. The drawback to this system is the complexity and the consuming training time for the BPNN. Furthermore, others have proposed the wavelet transform for feature extraction and vector quantization for the modelling technique, but, poor identification performance is achieved on the database used (15 speakers) [102]. Moreover, three scenarios for speaker identification were presented by [44], exploiting the GFM. However, the identification rate using the NIST 2003 database was poor. In addition, other researchers have examined large population speaker identification such as in [47] where the total variability space is used to capture both the speaker and channel variabilities by the I-vector. Their main challenge was providing a suitable database, so 1000 speakers were taken from a non standard YouTube database. The system gave higher performance for large observation periods (20s and more), but it was less efficient for shorter 10s speech lengths. In another study, a large population was achieved by using fuzzy clustering presented in [60], which employed hierarchical tree decisions for speaker identification. The study involved 3,805 speakers subjected to AWGN, and it was also noted that the system could be improved using fusion; however, no tests for realistic noise were conducted. In

## 4.1 Background

---

addition, In [103], a mean clustering approach was proposed for GMM speaker models, but the time complexity of the log-likelihood calculation was a bottleneck for the testing phase. The system achieved highest performance with TIMIT, with 10% and 30% reductions for the NIST 2002, and NTIMIT databases, respectively. However, the system was not evaluated under different environmental noise conditions.

In this chapter, a better identification performance is achieved for large populations compared with other work in [1] due to the following points: The PNCC features are robust to different types of noise and can sometimes achieve better SIA compared with MFCC and PLP features even in a clean environment [14]. Therefore, combining PNCC and MFCC features will provide a robust performance for original speech recordings and noisy environments which will be addressed in the next chapter. In contrast with prior work the proposed system also has the potential to achieve enhancement in SIA by removing and reducing sensitivity due to the channel between the speaker and microphone together with handsets by using normalization methods, feature warping and CMVN [1]. Moreover, instead of modelling individual speakers with limited data only by a GMM as in a previous study, a GMM-UBM is used based on modelling strategy as in [3] utilizing all speakers' data to increase the number of mixtures and thereby enhance the identification rate. Furthermore, this chapter studies a number of late fusion methods that includes weighted sum, maximum and mean fusion of the combination of the features scores as methods to improve SIA [76] [101]. In addition, early fusion and combination of early and late fusion are used [76]. Moreover, speaker identification with late fusion for static and dynamic features is also included in this work by using the vertical concatenated fusion methods between MFCC and PNCC features. However, a new method of speaker identification system is accomplished based on creation of a new score vector from scores vectors used in late, early and late fusion, and the concatenated of static and dynamic features mentioned above. This new vector was achieved via assuming that all fusion scores vectors are statistically independent which are essentially acquired from different feature dimensions 16, 32 and 39.

The organization for this chapter is as follows: Section 4.2 focus on biometric speaker identification framework. Fusion strategies are overviewed in Section 4.3. Section 4.4 includes the simulation setup. Overview of the related work is

explained in Section 4.5. Section 4.6 shows all original speech recordings simulation results for this chapter, while the discussions follow in Section 4.7. Finally, a summary is presented in Section 4.8 for this chapter.

## 4.2 Biometric Speaker Identification Framework

### 4.2.1 Feature Extraction and Feature Normalization

The major issue in the feature extraction is to transform the speech signal to compressed features that can provide a compact representation of the acoustic speech signal. Each speech sample in both training and testing phases has 129250 sample length (8 seconds train / 8 seconds test) in order to create a fixed length for all speech recordings which corresponds to a thousand frames for MFCC and PNCC features. Two feature extraction methods are combined MFCC and PNCC: a 16-feature dimension was selected from both MFCC and PNCC for each input frame including the average long-power (features at zero-order)  $C_0$  and  $P_{C_0}$  for both MFCC and PNCC respectively [3] [76] and [101]. The speech samples were filtered with a pre-emphasis filter by using a first order FIR high pass filter with emphasis coefficient 0.96 [1] [12]. Hamming windowing was exploited for both MFCC and PNCC features with frame duration 16ms with 50% inter frame rate overlap [1] [13]. The implementation for MFCC and PNCC features are explained in chapter 2, more details about both features can be found in [16] [26] [27] [104] [105].

Feature normalization is also adopted by using feature warping and CMVN for both the MFCC and PNCC features. The main purpose of feature warping is to produce a stronger representation of the distribution for each cepstral feature. For a specified time interval, warping the distribution of a stream of cepstral features to match the normal Gaussian is called feature warping [17]. Feature warping and CMVN approaches are used to improve the SIA for the system as well as reducing sensitivity to the mismatch between types of telephone handsets and could also help to reduce the linear channel effects [18] [19]. The features and feature compensations are developed in [76] [101].

### 4.2.2 Acoustic Modelling and Matching

Modelling a set of a speaker classes is one of the most important stages of the recognition task. In GMMs, each speaker can be represented by a finite weighted mixture of multivariate Gaussian components defined by the mean and covariance parameters as in Equations (4.1) and (4.2) [1] [41] [106]:

$$p(\mathbf{x} | \lambda) = \sum_{i=1}^M \omega_i p_i(\mathbf{x}) \quad (4.1)$$

where:  $\omega_i$  is the  $i$ -th mixture weight, and

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\mathbf{\Sigma}_i|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right\} \quad (4.2)$$

where:  $\mathbf{x}$  is a  $D$ -dimensional random feature vector,  $M$  is the number of Gaussian Mixture Components. For each speaker model there is a parameter set  $\lambda = \{\omega_i, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}, i = 1, \dots, M$ ,  $\boldsymbol{\mu}_i$  and  $\mathbf{\Sigma}_i$  are respectively the mean and covariance parameters of the  $i$ -th component density and  $(.)^T$  denotes the transpose operator [40]. Diagonal covariance matrices are assumed in this work instead of full covariance matrices as in [3] [106], which is more efficient computationally and no reduction in SIA is found. In most practical applications of GMMs in speaker identification only limited training data are available, therefore we use a different approach to learn these models as next described [76] [101].

#### 4.2.2.1 Universal Background Model and GMM-UBM

Basically, the GMM is one of the early methods that was used for modelling paradigms in speaker recognition. This method suffered from two major problems and they are unseen and insufficient data, therefore a limitation for increasing the Gaussian mixture dimension appeared and this caused degradation in the speaker identification system when the number of speakers increased. The GMM-UBM was proposed to solve the drawbacks from using GMM [3] and this is exploited in this chapter. One of the most important reasons to use the UBM is to overcome the problem of insufficient training data as well as unseen data. For understanding the UBM, imagine the UBM as a large pool in which all speaker training sets are used together with the Expectation Maximization (EM) method. As a consequence of training with a large amount of data a larger number of parameters are estimated than possible with individual speaker GMMs, and thereby increase the

## 4.2 Biometric Speaker Identification Framework

---

dimensionality of mixtures to cover all speakers which will improve the system performance. In addition, the individual speaker models are trained by Maximum A-Posteriori (MAP) adaptation initialized by the UBM with the training data for each particular speaker. This approach by training on large data for the UBM followed by adapting it for  $S$  different speakers will increase the dimensionality of the models. Therefore this coupling between the UBM ( large training data) and individual speaker models (small amount of data) makes the GMM-UBM system better able to model unseen data through having estimates with sufficient parameters to increase the mixture dimensionality (number of mixtures) and this improves the identification rate. The UBM represents an effective model for all non hypothesised speakers (in practice all the training speakers), and is defined as  $p(X | \lambda_{hyp})$  where  $X$  is the corresponding  $T_{\mathcal{F}}$  feature vectors,  $X = [\mathbf{x}_1, \dots, \mathbf{x}_{T_{\mathcal{F}}}]$ , and  $\lambda_{hyp}$  is the corresponding parameters set. This speaker independent model can then be used to improve speaker identification performance [76] [101].

### 4.2.2.2 Adaptation of Speaker Models

The parameters of the speaker models are found from the old estimates from the UBM training and the training data of the individual speakers. To control the balance between the new and old estimates, adaptation coefficients are used for weights, means and variances which can be represented by  $\alpha_i^w, \alpha_i^m, \alpha_i^v$  respectively. The adaptation coefficients are used in the following [76] [101] [106]:

$$\hat{\omega}_i = [\alpha_i^w n_i / T_{\mathcal{F}} + (1 - \alpha_i^w) \omega_i] \gamma \quad (4.3)$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\mathbf{x}) + (1 - \alpha_i^m) \boldsymbol{\mu}_i \quad (4.4)$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \alpha_i^v E_i(\mathbf{x}^2) + (1 - \alpha_i^v)(\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2) - \hat{\boldsymbol{\mu}}_i^2 \quad (4.5)$$

where  $\gamma$  is the scale factor to assure all adapted mixture weights have a unity summation. The data dependent mixing coefficients are calculated as in [106] as:

$$\alpha_i^\rho = \frac{n_i}{n_i + r^\rho} \quad (4.6)$$

$$n_i = \sum_{t=1}^{T_{\mathcal{F}}} \mathbf{Pr}(i | \mathbf{x}_t) \quad (4.7)$$

$$\mathbf{Pr}(i | \mathbf{x}_t) = \frac{\omega_i \mathbf{p}_i(\mathbf{x}_t)}{\sum_{j=1}^M \omega_j \mathbf{p}_j(\mathbf{x}_t)} \quad (4.8)$$

$$E_i(\mathbf{x}) = \frac{1}{n_i} \sum_{t=1}^{T_{\mathcal{F}}} \mathbf{Pr}(i | \mathbf{x}_t) \mathbf{x}_t \quad (4.9)$$

$$E_i(\mathbf{x}^2) = \frac{1}{n_i} \sum_{t=1}^{T_{\mathcal{F}}} \mathbf{Pr}(i | \mathbf{x}_t) \mathbf{x}_t^2 \quad (4.10)$$

where  $r^\rho$  is a fixed relevance factor,  $i$  is the mixture in the UBM,  $T_{\mathcal{F}}$  is the number of feature vectors,  $\mathbf{Pr}(i | \mathbf{x}_t)$  is the probabilistic alignment of the training vectors in the UBM mixture components [41] [106]. In addition, the parameters and adaptation coefficients used in the chapter can be listed as follows: for the initial UBM training  $final_{iter} = 20$ ; whereas for the MAP Adaptation the relevance factor  $r^\rho = 10$ ,  $\rho \in \{m, \omega, v\}$ ; and  $N_{mix} \in \{8, 16, 32, 64, 128, 256, 512\}$ ;  $\alpha_i^\rho \in [0, 1]$ . where:  $N_{mix}$  is the number of Gaussian Components.  $final_{iter}$  is the number of EM iterations. More details of the parameters and how they are used in the adaptation of speaker models can be found in [101] and [106]. Depending on the counts of data  $n_i$ , if  $\alpha_i^\rho \simeq 0$  for a speaker, the estimate relies more on the old sufficient statistics (low probabilistic count), while,  $\alpha_i^\rho = 1$  relies only on the new trained parameters (high probabilistic count), whereas the relevance factor  $r^\rho$  is used as a control between the new and old parameters [41] [106].

### 4.2.2.3 Maximum Log-likelihood Scores

Matching between training and testing is carried out by LLR. According to the Bayesian adaptation learning formula to apply the MAP adaptation; maximum log-likelihood should be achieved. The maximum a posteriori probability can be determined using equation (4.11) [41]:

$$S = \arg \max_{1 \leq k \leq S} P_r(\lambda_k | \mathbf{x}) = \arg \max_{1 \leq k \leq S} \frac{p(\mathbf{x} | \lambda_k) P_r(\lambda_k)}{p(\mathbf{x})} \quad (4.11)$$

where:  $S$  are a set of speakers,  $S = \{1, 2, \dots, S\}$ , which are represented by the GMM's models  $\lambda_1, \lambda_2, \dots, \lambda_S$ . The second part of (4.11) is because the Bayes' rule.

## 4.2 Biometric Speaker Identification Framework

---

Then, maximum likelihood classification can be derived as in equation (4.12) by assuming equally likely speakers with  $P_r(\lambda_k) = \frac{1}{S}$  and for all speaker models,  $p(\mathbf{x})$  is the same. The maximum likelihood can be determined as in (4.12) [41]:

$$S = \arg \max_{1 \leq k \leq S} p(\mathbf{x} | \lambda_k) \quad (4.12)$$

There are two reasons to use the Log function in the matching of likelihood one of them is that the log-function has monotonically increasing property that makes the maximum position unchanged after taking the log-likelihood for Gaussian models as well as the log function will cancel the exponential function the Gaussian of the GMM. The major purpose from ML estimation is to compute the speaker model parameters that can be able to maximize the likelihood of the GMM. For a non-linearity function such as GMM, it is not possible for direct maximization, therefore estimation ML can be done iteratively by using the EM algorithm. In this technique, initialization is started by choosing the initial model, then an expectation step is performed which probabilistically aligns vectors to generate a new model; while the maximization step is achieved by updating model parameters such that they be larger or equal to the initial model and then repeated until convergence is reached [106]. From a test speech signal (unknown speaker), features are extracted which form the inputs to the speaker models i.e. all speaker models  $S = \{\lambda_1, \lambda_2, \dots, \lambda_S\}$ . The log-likelihood scores are taken for the GMM-UBM system for each trial, which form a two-dimensional array with model-test set with a length 57,600 to represent the multiplication between 120 models with 480 tests (4 testing files for each speaker out of 120 speakers). So the trials are represented by model-test sets such as (Model 1, Test 1), ..., (Model 120, Test 1) to describe the scoring between all speaker models against the first test. However, each speaker has four tests therefore this will produce (Model 1, Test 2) ..., (Model 120, Test 2) and so on for 480 tests such as (Model 1, Test 480) to the (Model 120, Test 480) as shown in Fig. 4.1.



<b>Trials</b>	<b>Speaker Model</b>	<b>Test</b>
<b>1</b>	<b>1</b>	<b>1</b>
<b>2</b>	<b>2</b>	<b>1</b>
<b>3</b>	<b>3</b>	<b>1</b>
.	.	.
.	.	.
.	.	.
<b>120</b>	<b>120</b>	<b>1</b>
<b>121</b>	<b>1</b>	<b>2</b>
<b>122</b>	<b>2</b>	<b>2</b>
<b>123</b>	<b>3</b>	<b>2</b>
.	.	.
.	.	.
.	.	.
<b>240</b>	<b>120</b>	<b>2</b>
<b>241</b>	<b>1</b>	<b>3</b>
<b>242</b>	<b>2</b>	<b>3</b>
<b>243</b>	<b>3</b>	<b>3</b>
.	.	.
.	.	.
.	.	.
<b>360</b>	<b>120</b>	<b>3</b>
.	.	.
.	.	.
.	.	.
<b>57481</b>	<b>1</b>	<b>480</b>
<b>57482</b>	<b>2</b>	<b>480</b>
<b>57483</b>	<b>3</b>	<b>480</b>
.	.	.
.	.	.
.	.	.
<b>57600</b>	<b>120</b>	<b>480</b>

Figure 4.1: Trials Production for 120 Speakers From TIMIT Database with 120 speaker model and Four Testing Utterances Per Speaker (Total 480 Testing) to Yield 57,600 Trials of Model-Test Sets

The log-likelihood scores are calculated as [106] [101] [76].

$$LLR(X) = \log_e p(X | \lambda_{GMM}) - \log_e p(X | \lambda_{UBM}) \quad (4.13)$$

In this work four combination vectors of Log-likelihood scores are produced based on normalization and feature types. Each Log-likelihood score vector has a length of 57,600 scores as above, to represent 120 speakers by the scoring between 120 training speech files against 480 test files (4 speech samples for each speaker). In training and testing the resulting speaker models are scored with a maximum likelihood approach. For the final decision to identify each speaker the maximum log likelihood approach is used for speaker identification by taking the maximum score for each set of test scores for each speaker model, as [3] [41] [42]:

The system performance can be measured by the SIA which can be represented as in equation (3.4) [1] [100].

## 4.3 Speaker Identification Systems With Fusion Strategies

### 4.3.1 System 1: Speaker Identification System With Late Fusion

Fig. 4.2 represents the flowchart for late fusion speaker identification system. This system is based on four different combinations of normalized MFCC and PNCC features to represent multi bases with and without late fusion (late fusion is score based fusion including three methods maximum, mean and weighted sum) for 16 feature dimension. Point A denotes scores for the normalized MFCC features such as FWMFCC or CMVNMFCC, similarly point B represents the scores for the normalized PNCC features (FWPNCC or CMVNPNC). Four identification systems (before fusion) can be produced by connecting either point A or B to the point X, the first time for Feature Warping, while the second for CMVN features. Two identification systems are produced by connecting point A with X depending either upon applying FW or CMVN to the MFCC features and similarly for

### 4.3 Speaker Identification Systems With Fusion Strategies

PNCC features at point B. On the other hand, late fusion is developed by connecting Points A, B to the points S1, S2 respectively using one of the fusion methods: maximum, mean and weighted sum [76] [101]. Three late fusion of scores approaches are adopted: Depending on the features and normalization methods, four combinations of log-likelihood scores are constructed. These are:  $\mathbf{f}_1$ = Feature warping MFCC scores vector (FWMFCC),  $\mathbf{f}_2$ = CMVN MFCC scores vector,  $\mathbf{g}_1$ = Feature Warping PNCC scores vector (FWPNCC) and  $\mathbf{g}_2$ = CMVN PNCC scores vectors. These score vectors are found before the fusion process and form the following composite vectors as (4.14a)and (4.14b).

$$\mathbf{f}_i = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{bmatrix} \quad (4.14a)$$

$$\mathbf{g}_j = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} \quad (4.14b)$$

To produce four types of maximum of fusion scores vectors:  $\mathbf{fmax}_{11}$ ,  $\mathbf{fmax}_{12}$ ,  $\mathbf{fmax}_{21}$  and  $\mathbf{fmax}_{22}$ , then the row wise maximum is evaluated as in equation (4.15).

$$\mathbf{fmax}_{ij} = \max(\mathbf{f}_i, \mathbf{g}_j) \quad (4.15)$$

where:  $\mathbf{fmax}_{ij}$  is the fusion maximum scores vector, i, j = 1, 2.

Similarly, to produce four mean fusion scores vectors:  $\mathbf{fmean}_{11}$ ,  $\mathbf{fmean}_{12}$ ,  $\mathbf{fmean}_{21}$  and  $\mathbf{fmean}_{22}$ , then equation (4.16) is used to calculate the fusion mean.

$$\mathbf{fmean}_{ij} = (\mathbf{f}_i + \mathbf{g}_j)/2 \quad (4.16)$$

where:  $\mathbf{fmean}_{ij}$  is the fusion mean scores vector, i, j = 1, 2.

In addition, a linear weighted sum fusion of scores is used for the scores vectors:  $\mathbf{fweight}_{11}$ ,  $\mathbf{fweight}_{12}$ ,  $\mathbf{fweight}_{21}$  and  $\mathbf{fweight}_{22}$ , as in equation (4.17).

$$\mathbf{fweight}_{ij} = \omega_\beta \mathbf{f}_i + (1 - \omega_\beta) \mathbf{g}_j \quad (4.17)$$

where:  $\beta = 1, 2, 3, 4$ . while,  $\omega_1, \omega_2, \omega_3$  and  $\omega_4 = 0.9, 0.8, 0.77$  and  $0.7$  respectively. where, both i and j take values 1 and 2, therefore  $\mathbf{fweight}_{ij}$  takes one of four values  $\mathbf{fweight}_{11}$ ,  $\mathbf{fweight}_{12}$ ,  $\mathbf{fweight}_{13}$  and  $\mathbf{fweight}_{22}$ , and  $\mathbf{fweight}_{11}$  is the linear combination of  $\mathbf{f}_1$  and  $\mathbf{g}_1$ , likewise  $\mathbf{fweight}_{12}$  is the linear

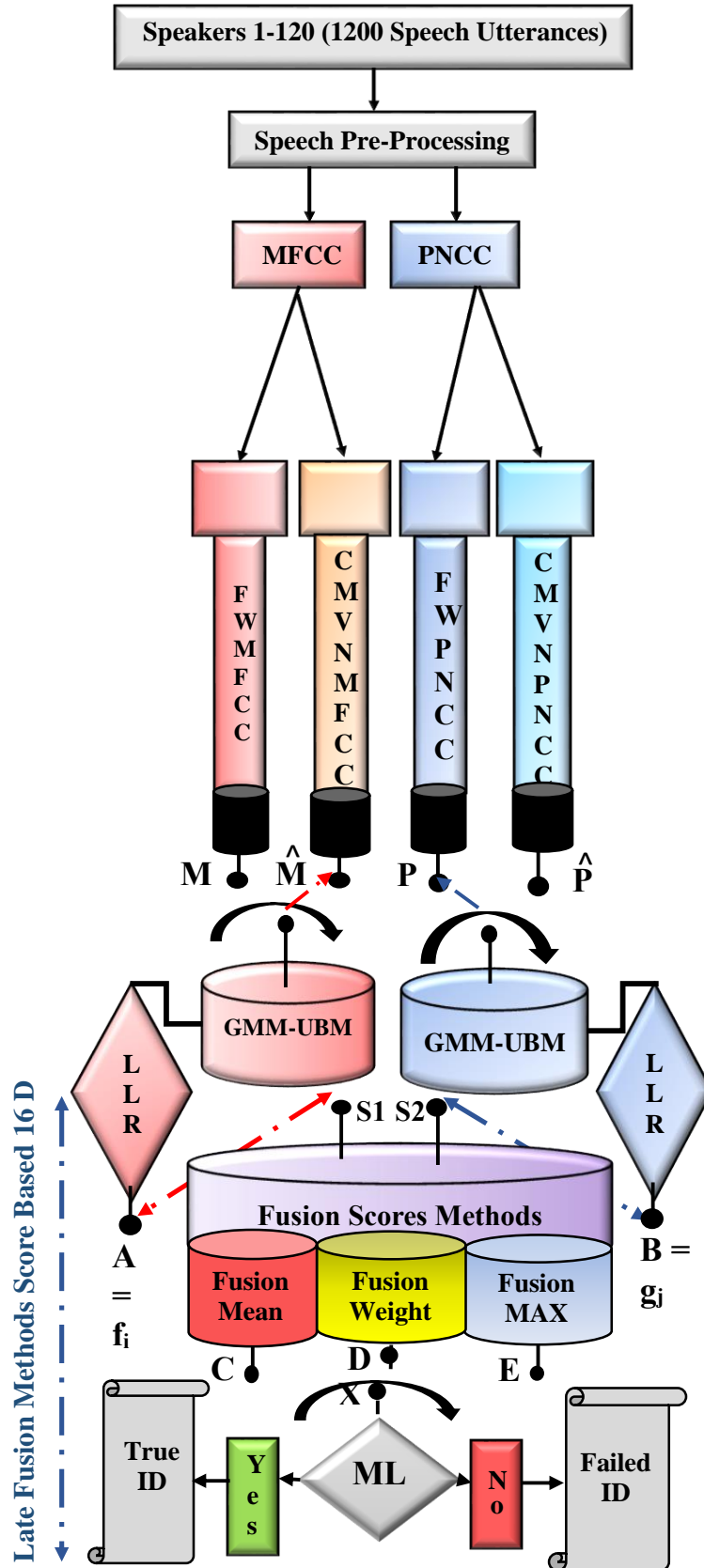


Figure 4.2: Flowchart for Speaker Identification System Multi-Bases (16D) With/Without (W/O) Late Fusion

combination of  $\mathbf{f}_1$  and  $\mathbf{g}_2$  and so on. For each  $\mathbf{fweight}_{ij}$ ,  $\omega_\beta$  can take on one of four values namely,  $\omega_\beta \in \{0.9, 0.8, 0.77, 0.7\}$  which is chosen to give empirically the best SIA.  $\omega_\beta$  is limited to these four values as lower values have been found to be unsuitable to yield high SIA performance, because MFCC coefficients are more important in the speaker identification task with clean speech.

#### 4.3.2 System 2: Speaker Identification System With Early Fusion and Early-Late Fusion

In Fig. 4.3 four bases systems are formed by vertical concatenation from normalized 16 MFCC features (either FWMFCC or CMVNMFCC) with the corresponding 16 normalized PNCC features (either FWPNCC or CMVNPCC) to produce 32 features dimension which represent the early fusion. These early systems can be modified using late fusion methods to improve the SIA [76]. In early fusion the main feature fusion process is performed before modelling the system by GMM-UBM and this process is limited by the same fixed size for each feature type. The system can be described by the following [76]:  $\mathbf{M} = \{\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_k\}$ ,  $\hat{\mathbf{M}} = \{\hat{\mathbf{M}}_1, \hat{\mathbf{M}}_2, \dots, \hat{\mathbf{M}}_k\}$ ,  $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_k\}$ ,  $\hat{\mathbf{P}} = \{\hat{\mathbf{P}}_1, \hat{\mathbf{P}}_2, \dots, \hat{\mathbf{P}}_k\}$ , where:  $\mathbf{M}$  = FWMFCC feature matrices,  $\hat{\mathbf{M}}$  = CMVNMFCC feature matrices,  $\mathbf{P}$  = FWPNCC feature matrices,  $\hat{\mathbf{P}}$  = CMVNPCC feature matrices,  $k$  is equal 720 for training phase and 480 for testing side, each feature matrix =  $D \times N$ ,  $D$ = feature dimension is equal 16,  $N$ = numbers of frames is equal 1000. Equation (4.18) (can be used to explain early feature fusion as :

$$\mathbf{H}_{i,j} = \begin{bmatrix} \mathbf{MFCC}_i & \mathbf{PNCC}_j \end{bmatrix}, \quad i, j = 1, 2. \quad (4.18)$$

where:  $\mathbf{MFCC}_i$  is normalized MFCC feature matrices before modelling which contains either  $\mathbf{MFCC}_1$  as  $\mathbf{M}$ , or  $\mathbf{MFCC}_2$  as  $\hat{\mathbf{M}}$  likewise for PNCC,  $\mathbf{PNCC}_i$  = is normalized PNCC feature matrices before modelling where,  $\mathbf{PNCC}_1$  is  $\mathbf{P}$  and  $\mathbf{PNCC}_2$  is  $\hat{\mathbf{P}}$ . Similarly, the system after matching can identify four bases lines of early feature fusion by connecting Points A, B by the maximum likelihood (ML) by Point X as well the system can extend to include the late fusion to produce early-late fusion. This can be achieved by connecting points A, B to the points S1, S2, respectively, then using one of the late fusion methods.

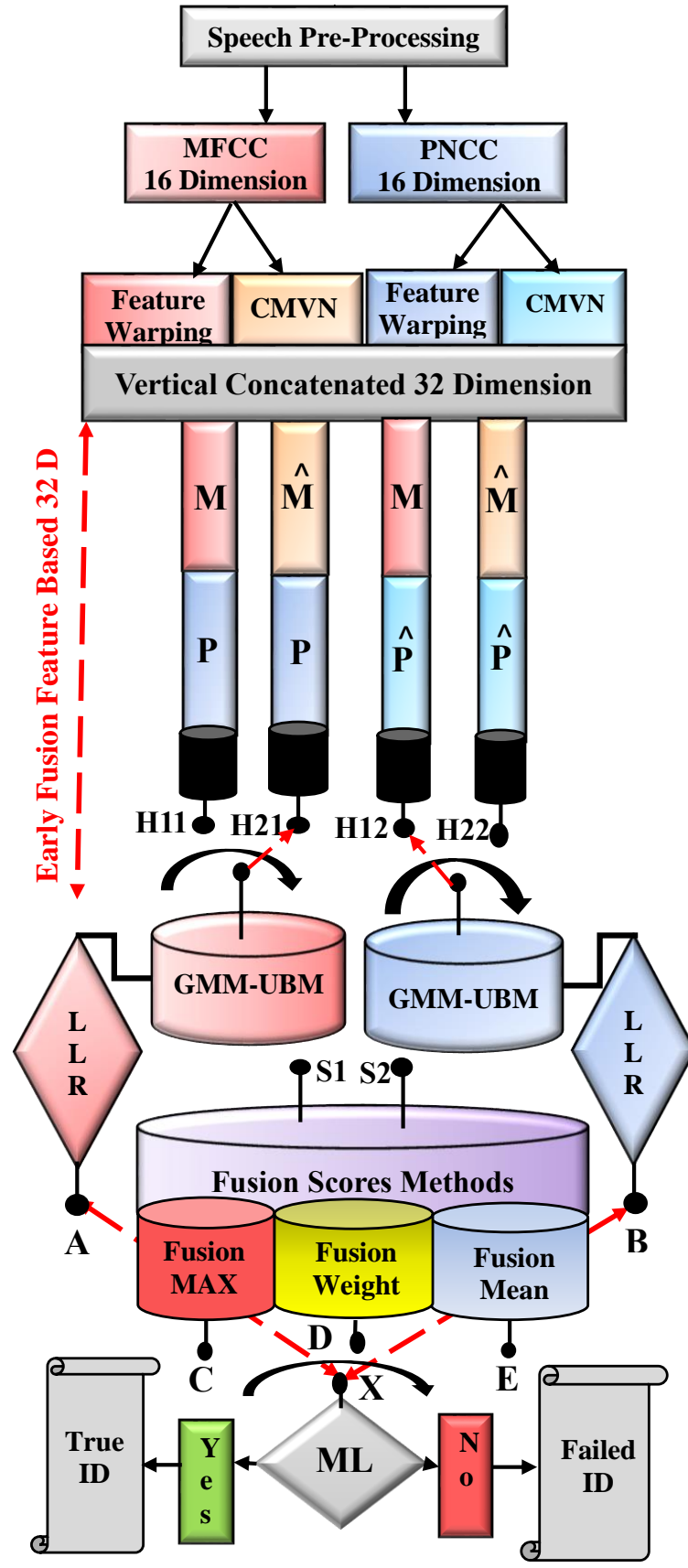


Figure 4.3: Flowchart for Speaker Identification System Multi-Bases (32D) Early Fusion With/Without Late Fusion

#### 4.3.3 System 3: Speaker Identification System With Late Fusion for Concatenated Static and Dynamic Features

This system is quite similar to the system 1, while the main difference in this system is the extension in the feature dimension to the feature parametrization to include the static and dynamic MFCC and PNCC features to 39 features per frame. This is achieved by concatenate 13 MFCC/PNCC features (original features) with temporal derivatives including 13 features from the the first order derivative (Delta) as well as the corresponding 13 features from the second order derivatives (Delta-Delta) to yield 39 features such as  $FeatureDim(39 MFCC) = 13 MFCC + 13 \Delta MFCC + 13 \Delta\Delta MFCC$ ,

likewise for PNCC features

$FeatureDim(39 PNCC) = 13 MFCC + 13 \Delta PNCC + 13 \Delta\Delta PNCC$  [3]. Late fusion methods score based are applied to the normalized methods for MFCC (FWMFCC and CMVNMFCC) with the corresponding scores for PNCC normalized features (FWPNCC and CMVNPNC). Equation (4.19) is applied to calculate the delta feature (first order derivative) for both MFCC and PNCC [3].

$$\mathbf{d}_t = \sum_{\delta=1}^{del} \frac{\delta(\mathbf{c}_{t+\delta} - \mathbf{c}_{t-\delta})}{2 \sum_{\delta=1}^{del} \delta^2} \quad (4.19)$$

where:  $del$  is typically 2,  $\mathbf{c}_t = [c_0, c_1, \dots, c_L]$ ,  $L = 12$ ,  $\mathbf{c}$  is either MFCC or PNCC features which is 13 coefficients ( $c_0+L$ ),  $\mathbf{d}_t$  is the feature vector for the first order derivative,  $t$  is frame time index. Similarly, the acceleration parameter  $\mathbf{a}_t$  vectors which represent the second order derivative vectors can be produced by replacing  $\mathbf{c}_t$  with the  $\mathbf{d}_t$  in equation (4.16). Therefore in order to boost the original MFCC and PNCC features, temporal derivatives  $\mathbf{d}_t$  and  $\mathbf{a}_t$  as dynamic features are concatenated with the static features  $\mathbf{c}_t$ . This strategy was adopted for MFCC only in speech recognition as well in recent researches in speaker recognition [3] as explained in equation (4.20) as illustrated in the Fig. 4.4 . In this chapter, this technique is exploited not only for MFCC features, but also for PNCC features.

$$\mathbf{cf}_t = \begin{bmatrix} \mathbf{c}_t^T & \mathbf{d}_t^T & \mathbf{a}_t^T \end{bmatrix}^T \quad (4.20)$$

### 4.3 Speaker Identification Systems With Fusion Strategies

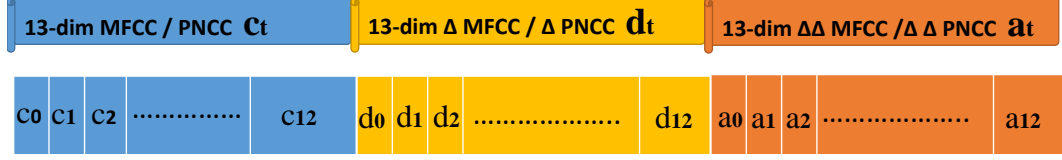


Figure 4.4: Concatenated MFCC/PNCC Static Features with the Dynamic Features with 39 Dimension [3]

where:  $\mathbf{cf}_t$  is the concatenated the static and dynamic temporal features for (MFCC/ PNCC) with 39 dim.

#### 4.3.4 System 4: Speaker Identification System With Late Fusion For Normalized Independent Scores For Systems 1, 2 and 3

This system was deduced from the vector scores from the three systems above; these scores vectors are normalized for constancy of scores by subtracted each vector from the mean for that vector and the result is divided by the standard deviation for this vector. These scores are assumed statistically independent where the dimension for each scores vector is 57600, while these scores are coming from different feature dimensions (16, 32 and 39). The main purpose is to accomplish a new scores vectors which are created by the multiplication of three independent scores vectors that essentially developed from different features dimension (16, 32 and 39). This new system is quite similar to those in genetic development systems where the new system has scores are originally comes from mixing three different features dimensions. The scores vectors for system 1, system 2 and system 3 are normalized and the equations in (4.21), (4.22) and (4.23) are adopted to present system 4.

$$\mathbf{sysA}_i = \frac{\mathbf{sysa}_i - \mu(\mathbf{sysa}_i)}{\sigma(\mathbf{sysa}_i)} \quad (4.21)$$

$$\mathbf{sysB}_i = \frac{\mathbf{sysb}_i - \mu(\mathbf{sysb}_i)}{\sigma(\mathbf{sysb}_i)} \quad (4.22)$$

$$\mathbf{sysC}_i = \frac{\mathbf{sysc}_i - \mu(\mathbf{sysc}_i)}{\sigma(\mathbf{sysc}_i)} \quad (4.23)$$



#### 4.4 Simulations Setup

---

where:  $\mathbf{sysa}_i$ ,  $\mathbf{sysb}_i$  and  $\mathbf{sysc}_i$  represent the four scores vectors for system 1, system 2 and system 3 respectively.  $\mathbf{sysA}_i$ ,  $\mathbf{sysB}_i$  and  $\mathbf{sysC}_i$  are the normalized scores vectors for previous systems respectively.  $i = 1, 2, 3, 4$ .  $\mathbf{sysA}_i$  are the normalized scores vectors used to the scores vectors of system 1 as the following:  $\mathbf{f}_1, \mathbf{f}_2, \mathbf{g}_1, \mathbf{g}_2$  for features 16 dim.  $\mathbf{sysB}_i$  are the normalized scores vectors used to the system 2 as the following:  $\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \mathbf{h}_4$  for features 32 dim. Finally  $\mathbf{sysC}_i$  are the normalized scores vectors used to the system 3 as the following:  $\mathbf{\ddot{f}}_1, \mathbf{\ddot{f}}_2, \mathbf{\ddot{g}}_1, \mathbf{\ddot{g}}_2$  for features 39 dim. The new system is assumed all the normalized scores for the systems 1, 2 and 3 ( $\mathbf{sysA}_i$ ,  $\mathbf{sysB}_i$  and  $\mathbf{sysC}_i$ ) are statistically independent thereby multiplications for these scores vectors are applied as in (4.24).

$$\mathbf{sysd}_i = \mathbf{sysA}_i \cdot \mathbf{sysB}_i \cdot \mathbf{sysC}_i \quad (4.24)$$

where:  $(\cdot)$  represent the element wise multiplication,  $i=1, 2, 3$  and 4.

$$\mathbf{sysd}_1 = \mathbf{f}_1 \cdot \mathbf{h}_1 \cdot \mathbf{\ddot{f}}_1$$

$$\mathbf{sysd}_2 = \mathbf{f}_2 \cdot \mathbf{h}_2 \cdot \mathbf{\ddot{f}}_2$$

$$\mathbf{sysd}_3 = \mathbf{g}_1 \cdot \mathbf{h}_3 \cdot \mathbf{\ddot{g}}_1$$

$$\mathbf{sysd}_4 = \mathbf{g}_2 \cdot \mathbf{h}_4 \cdot \mathbf{\ddot{g}}_2$$

There are four scores vectors are produced as a consequent of applying (4.24) and similarly these vectors are normalized and (4.25) is used for normalization purpose as explained in Fig. 4.5.

$$\mathbf{sysd(norm)}_i = \frac{\mathbf{sysd}_i - \mu(\mathbf{sysd}_i)}{\sigma(\mathbf{sysd}_i)} \quad (4.25)$$

#### 4.4 Simulations Setup

In all the simulations, training and testing are conducted on a personal computer with Intel(R) Core(TM) i5-3470 CPU 3.20 GHz with Installed Memory (RAM) 16.0 GB, and Windows 7 copyright©2009 service pack1 as an operating system. The speech sampling frequency used is 16 kHz. In all the experiments for this chapter, the TIMIT database was employed because it is a common speech corpus, widely available and exploited in [3], [76] and [101]. 49 speakers are selected from Dialect Region one (DR 1) and 71 speakers from (DR 4) to mirror those used in [1] to

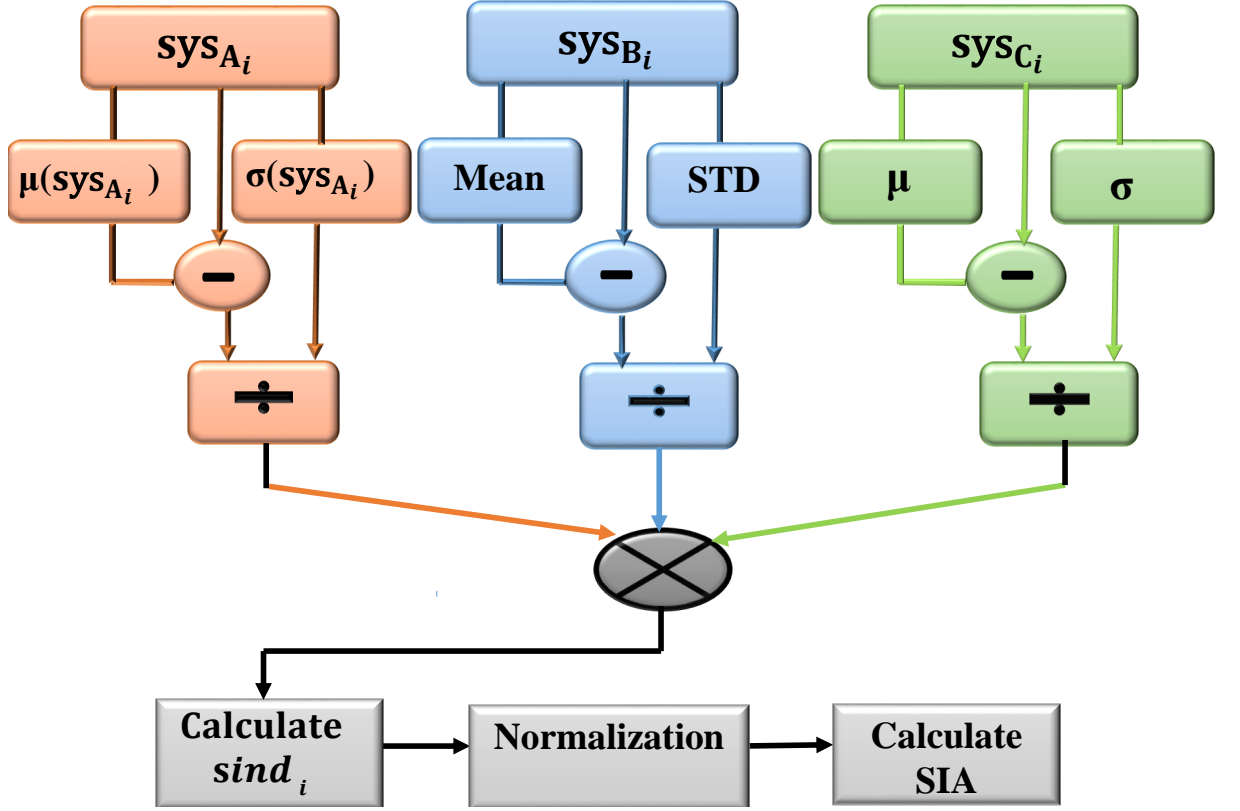


Figure 4.5: System 4 Independent Scores for Different Feature Dimensions

produce totally 120 speakers (1200 speech samples). All speech samples were taken with fixed length of 129250 samples; concatenation is applied where necessary. There are eight dialects in the TIMIT database, in this chapter both DR 1 from New England and DR 4 from South Midland are employed. The main parameters used in the experiments are explained in Table 4.1, these parameters can be categorized according to database, dialect region, sampling rate, window size and frame shift, pre-emphasis factor, window type, number of speakers used, number of samples per speaker, total number of samples used, training and testing samples per speaker and finally the average duration for each speech sample [76] [101].

## 4.5 Related Work

Table 4.2 illustrates a comparison among all aspects between the work in [1] and all other simulations in this chapter. In addition, Table 4.2 gives a summary of all simulations infrastructures for this chapter compared with [1] such as speakers dialect region, features and feature dimension used, feature compensation methods (normalization), modelling, Gaussian mixture components (GMCs), type of

## 4.6 Simulation Results

Table 4.1: Experimental Parameters for the Work in [1] and in All Proposed Simulations in This Chapter

Database	TIMIT
Dialect	DR1 and DR4
Sampling frequency	16000
Window size	16 ms
Frame shift	8 ms
Pre-emphasis factor	0.96
Window type	Hamming
Number of speakers	120
No. of samples per speaker	10
Total samples used	1,200
Training samples	6 per speaker(total 720)
Testing samples	4 per speaker(total 480)
Average sample duration	8 seconds (work [1] 3 Seconds)

classifier, fusion types, speaker identification accuracy and the system environment. The work in [1] used 120 speakers from two dialect regions DR1 and DR4 which are randomly chosen without mentioning how many speakers are taken from each dialect region, whereas all speakers (49 speakers) are taken from DR1 and 71 speakers from DR4. Furthermore, in [1] MEL is used without using any kind of normalization. In this work two types of features are employed one of them is robust for noise (PNCC) which is fused with MFCC features which are efficient for original speech recordings; in addition feature normalization is investigated to solve the linear channel effect problems. Table 4.3 shows results for the approach in [1] with three settings of the number of Gaussian mixtures components in [1], {8, 16, 32}. The highest SIA achieved in Table 4.3 is 93.88% at the mixture size 16 whilst continuous increasing the mixture components caused decreasing the SIA, the reason for that is for limited GMM data trained and this is one of the most important problems are tackled in this work. Furthermore, in [1] only Mel features is used with and fused with Inverse MEL features (IMEL) by using the fusion weights: 0.5, 0.7, 0.77 and 0.8.

## 4.6 Simulation Results

### 4.6.1 Simulation Results For System 1

According to Table 4.4, presents the simulation results for system 1 which include the SIA based on late fusion. This table shows the SIA to the four combinations

## 4.6 Simulation Results

Table 4.2: Main Comparison Between the Work in [1] and the Proposed Algorithm

Aspects	Methods in [1]	The Proposed Simulations in This Chapter
Speakers DR	DR1& DR4	49 DR1& 71 DR4
Features	1-MEL 2-IMEL	1-MFCC 2-PNCC
Features Dim	N/A	16
Feature Norm.	Not Used	FW and CMVN
Modelling	GMM	GMM-UBM
GMCs (Mixtures)	[8, 16, 32]	[8, 16, 32, 64, 128, 256, 512]
Classifier	LLR	LLR
Fusion Types	Fusion Weight	Early, Late, Early-Late, Concatenated Statistic and Dynamic features and Score independent fusion
SIA	93.88%	<b>95%</b>
System Environment	Clean	Clean and Noisy (Noisy in the next chapter)

Table 4.3: SIA Results for Original Speech Recordings as in [1]

SIA Results for Work in [1]			
Methods	Mix8	Mix16	Mix32
MEL	67.35%	74.36%	71.43%
IMEL	55.10%	58.97%	53.06%
Fused $\omega_1=0.5$	79.59%	87.75%	83.67%
Fused $\omega_2=0.7$	88.2%	90.31%	89.15%
Fused $\omega_3=0.77$	89.8%	<b>93.88%</b>	91.84%
Fused $\omega_4=0.8$	89.8%	91.84%	91.84%

of features based on MFCC and PNCC features for different GMCs namely {8, 16, 32, 64, 128, 256, 512}. These combinations are: FWMFCC, CMVNMFCC, FWPNCC and CMVNPNC. It is clear from the Table 4.4 that the SIA for the MFCC features is higher as compared with the corresponding results in PNCC features and that because the MFCC have a better performance compared with PNCC in clean environments. The scores for the best SIA between the MFCC features (FWMFCC and FWMCC i.e ( $\mathbf{f}_1$ ) or ( $\mathbf{f}_2$ )) are fused with the corresponding scores for the best SIA for the PNCC features (FWPNCC and FWPNC i.e  $\mathbf{g}_1$  or  $\mathbf{g}_2$ ). Then three late fusion methods are applied to the scores vectors belongs to the fusion decision; the late fusion methods are weighted sum, maximum and mean fusion methods. This work presented four main weights; 0.9, 0.8, 0.77, 0.7, while the weight 0.77 is selected to mirror the work in [1] for comparison purpose. The first

## 4.6 Simulation Results

highest SIA is achieved by the fusion weights with weight 0.9 and at mixture size 512 with SIA 95%, whereas the second highest SIA is accomplished at the weighted sum 0.8, 0.77, 0.7 as well at the mean fusion at mixture sizes 512 and 256 with SIA 94.17%. However, related to all mixtures sizes the weighted sum fusion appears to be the best fusion methods compared with all late fusion methods. There is a benefit in SIA from fusion both MFCC and PNCC features together as compared with each feature alone. In addition, increasing the Gaussian mixture component size causes additional increment in the data trained by GMM-UBM then this will increase the accuracy for the SIA. [76].

Table 4.4: Simulation 1: Speaker Identification System with Late Fusion

Speaker Identification Accuracy (SIA %) for Different GMCs							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
FWMFCC ( $\mathbf{f}_1$ )	80%	84.17%	89.17%	93.33%	93.33%	93.33%	94.17%
CMVNMFCC ( $\mathbf{f}_2$ )	77.5%	80.83%	86.67%	91.67%	91.67%	92.5%	90.83%
FWPNCC ( $\mathbf{g}_1$ )	60%	71.67%	80.83%	86.67%	88.33%	90%	90%
CMVNPNC ( $\mathbf{g}_2$ )	70%	74.17%	83.33%	86.67%	90%	89.17%	90.83%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fused $\omega_1=0.9$	79.17%	84.17%	90%	91.67%	93.33%	93.33%	<b>95%</b>
Fused $\omega_2=0.8$	80.83%	83.33%	90%	92.5%	93.33%	94.17%	94.17%
Fused $\omega_3=0.77$	80.83%	83.33%	90%	92.5%	93.33%	94.17%	94.17%
Fused $\omega_4=0.7$	79.17%	82.5%	89.17%	92.5%	93.33%	94.17%	93.33%
Fusion Max	77.5%	75%	85%	90%	94.17%	93.33%	93.33%
Fusion Mean	78.33%	80.83%	90%	92.5%	93.33%	94.17%	94.17%

### 4.6.2 Simulation Results For System 2

Table 4.5, presents the simulation results for system 2 which is early fusion and early-late fusion. This table provides four early fusions by vertically concatenated 16 MFCC features (FWMFCC and CMVNMFCC) with the corresponding 16 PNCC features (FWPNCC and CMVNPNC) to produce four combinations each with 32 dimension and the early fused matrices with 32 feature dimensions before modelling are explained as the following:  $\mathbf{H}_{1,1}$ ,  $\mathbf{H}_{2,1}$ ,  $\mathbf{H}_{1,2}$  and  $\mathbf{H}_{2,2}$ . Early fusion scores vectors for the corresponding previous matrices are:  $\mathbf{h}_1$ ,  $\mathbf{h}_2$ ,  $\mathbf{h}_3$  and  $\mathbf{h}_4$ . where the highest SIA for the early fusion is accomplished at score vector  $\mathbf{h}_3$  with SIA 91.67% at mixture size 256. Late fusion strategies to the early scores vectors are used to improve the SIA belongs to the fusion decision which is effectively fusion the scores vector for highest SIA between  $\mathbf{h}_1$  and  $\mathbf{h}_3$  with the corresponding highest SIA for the scores vector between both  $\mathbf{h}_2$  and  $\mathbf{h}_4$ . Highest SIA accuracy is achieved via

## 4.6 Simulation Results

fusion  $\mathbf{h}_3$  with the  $\mathbf{h}_4$  92.5% at mixture size 256 at weighted sum fusion for all weights used and at fusion mean. On the other hand, the fusion mean is appears as a better performance accuracy than other fusion methods. There is an advantage in SIA from fusion and increasing the mixture sizes causes additional increment in the SIA, while increasing the GMCs size to the 512 will started to reduce the SIA. This mean, the best SIA can be achieved at mixture size 256.

Table 4.5: Simulation 2: Speaker Identification System with Early Fusion and Early-Late Fusion

Speaker Identification Accuracy (SIA %) for Different GMCs							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
$\mathbf{h}_1$ (Scores of $\mathbf{H}_{1,1}$ )	61.67%	66.67%	76.67%	86.67%	90.83%	89.17%	86.67%
$\mathbf{h}_2$ (Scores of $\mathbf{H}_{2,1}$ )	65%	65%	73.33%	84.17%	86.67%	88.33%	85.83%
$\mathbf{h}_3$ ( Scores of $\mathbf{H}_{1,2}$ )	64.17%	70.83%	75%	85%	90%	91.67%	86.67%
$\mathbf{h}_4$ ( Scores of $\mathbf{H}_{2,2}$ )	65%	65%	79.17%	85.83%	91.67%	90%	89.17%
Fusion Decision	$\mathbf{h}_3\text{-}\mathbf{h}_4$	$\mathbf{h}_3\text{-}\mathbf{h}_4$	$\mathbf{h}_1\text{-}\mathbf{h}_4$	$\mathbf{h}_1\text{-}\mathbf{h}_4$	$\mathbf{h}_1\text{-}\mathbf{h}_4$	$\mathbf{h}_3\text{-}\mathbf{h}_4$	$\mathbf{h}_3\text{-}\mathbf{h}_4$
Fused $\omega_1=0.9$	64.17%	70%	76.67%	87.5%	90.83%	<b>92.5%</b>	86.67%
Fused $\omega_2=0.8$	65.83%	71.67%	77.5%	88.33%	90.83%	<b>92.5%</b>	87.5%
Fused $\omega_3=0.77$	65.83%	71.67%	77.5%	88.33%	90.83%	<b>92.5%</b>	87.5%
Fused $\omega_4=0.7$	65%	71.67%	78.33%	89.17%	91.67%	<b>92.5%</b>	88.33%
Fusion Max	64.17%	66.67%	80%	88.33%	90%	90.83%	88.33%
Fusion Mean	65%	68.33%	79.17%	90.83%	90.83%	<b>92.5%</b>	89.17%

## 4.6 Simulation Results

### 4.6.3 Simulation Results For System 3

Table 4.6, provides the simulation results for system 3 which contained speaker identification system with late fusion for the concatenated of static and dynamic features. Generally, this table is similar to the Table 4.4; where is only one main difference is the feature dimension used is extended to the 39 features instead of 16 that used in Table 4.4. The concatenation are adopted for 13 features from MFCC, Delta and Delta Delta to produced 39 features and likewise for PNCC features. Then fusion decision between the highest SIA for the concatenated MFCC features ( $\ddot{f}_1, \ddot{f}_2$ ) with the corresponding PNCC features ( $\ddot{g}_1, \ddot{g}_2$ ). It is evident from the Table 4.6 that the mixture size and the dynamic features that represented by the first and second derivatives are playing the significant role for calculation the SIA. Moreover, it can recognize about 50% dropping in SIA at mixture size 8 as compared with Table 4.4 and Table 4.5, while acceptable SIA can be achieved by increasing the GMCs (128-512). The maximum SIA occurs at mixture size 512 with 87.5% at weights 0.9 and 0.8 as well as at fusion mean when fused  $\ddot{f}_2$  with  $\ddot{g}_2$ .

Table 4.6: Simulation 3: Speaker Identification System with Late Fusion for the Concatenated of the Static and Dynamic Features

Speaker Identification Accuracy (SIA %) for Different GMCs							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
<i>cf of</i> FWMFCC ( $\ddot{f}_1$ )	40.83%	44.17%	49.17%	66.67%	74.17%	82.5%	80%
<i>cf of</i> CMVNMFCC ( $\ddot{f}_2$ )	42.5%	47.5%	57.7%	69.17%	84.17%	85%	85.83%
<i>cf of</i> FWPNCC ( $\ddot{g}_1$ )	36.67%	40%	43.33%	60.83%	67.5%	71.67%	78.33%
<i>cf of</i> CMVNPNC ( $\ddot{g}_2$ )	38.33%	40%	48.33%	65%	80%	75%	80%
Fusion Decision	( $\ddot{f}_2-\ddot{g}_2$ )	( $\ddot{f}_2-\ddot{g}_2$ )	( $\ddot{f}_2-\ddot{g}_2$ )	( $\ddot{f}_2-\ddot{g}_2$ )	( $\ddot{f}_2-\ddot{g}_2$ )	( $\ddot{f}_2-\ddot{g}_2$ )	( $\ddot{f}_2-\ddot{g}_2$ )
Fused $\omega_1=0.9$	45.83%	48.33%	61.67%	71.67%	85%	87.5%	<b>87.5%</b>
Fused $\omega_2=0.8$	45%	50%	63.33%	72.5%	86.67%	86.67%	<b>87.5%</b>
Fused $\omega_3=0.77$	45%	51.67%	63.33%	72.5%	85%	86.67%	86.67%
Fused $\omega_4=0.7$	45%	50.83%	63.33%	73.33%	85%	86.67%	86.67%
Fusion Max	41.67%	43.33%	53.33%	65%	79.17%	82.5%	85%
Fusion Mean	47.5%	50.83%	58.33%	72.5%	84.17%	86.67%	<b>87.5%</b>

### 4.6.4 Simulation Results For System 4

Table 4.7, gives the simulation results for system 4 which contained the fusion for normalized independent scores for systems 1, 2 and 3. Similarly, Table 4.4, Table 4.5 and Table 4.6 are constructed. Each scores vector in  $\mathbf{ind(norm)}_i$  is developed from the normalized to the element multiplications for statistically independent

## 4.6 Simulation Results

vectors in three previous simulations. The highest SIA achieved at mixture size 128 at fusion maximum between  $\mathbf{ind(norm)}_2$  and  $\mathbf{ind(norm)}_4$  with SIA 93.33%. Generally, the results achieved by this table are better than those in Table 4.5 and Table 4.6. The most important issue is that the scores used are created from different scores essentially from three different features dimensions for instance 16, 32 and 39. Therefore, similar to genetic development it can improved the performance by fusion scores with different features dimensions.

Table 4.7: Simulation 4: Speaker Identification System with Late Fusion for Normalized Independent Scores for Systems 1, 2 and 3

Speaker Identification Accuracy (SIA %) for Different GMCs							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
$\mathbf{ind(norm)}_1 = 1$	79.17%	85%	84.17%	89.17%	91.67%	90.83%	92.5%
$\mathbf{ind(norm)}_2 = 2$	78.33%	83.33%	86.67%	90%	91.67%	91.67%	90.83%
$\mathbf{ind(norm)}_3 = 3$	68.33%	70.83%	76.67%	83.33%	85.83%	85.83%	86.67%
$\mathbf{ind(norm)}_4 = 4$	70.83%	71.67%	79.17%	85.83%	88.33%	87.5%	89.17%
Fusion Decision	(1,4)	(1,4)	(2-4)	(2,4)	(2-4)	(2-4)	(1-4)
Fused $\omega_1=0.9$	78.33%	83.33%	86.67%	89.17%	91.67%	91.67%	90.83%
Fused $\omega_2=0.8$	79.17%	83.33%	85.83%	89.17%	91.67%	91.67%	90.83%
Fused $\omega_3=0.77$	79.17%	81.67%	85.83%	89.17%	91.67%	91.67%	90%
Fused $\omega_4=0.7$	79.17%	80.83%	85.83%	89.17%	91.67%	91.67%	90%
Fusion Max	77.5%	78.33%	84.17%	90.83%	<b>93.33%</b>	92.5%	92.5%
Fusion Mean	77.5%	78.33%	85%	89.17%	90.83%	91.67%	90.83%



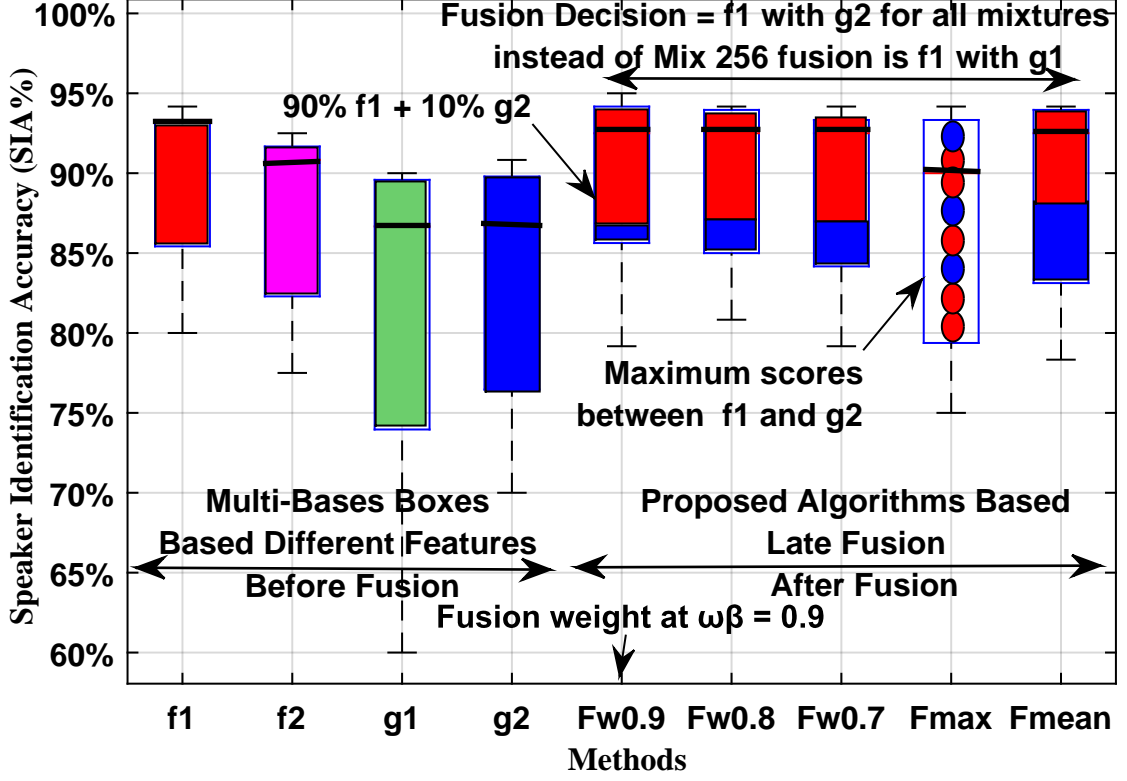


Figure 4.6: Box Plot of SIA for Original Speech Recordings with Multi-Bases and Late Fusion Proposed Algorithms for 16D and Relate to Simulation 1, with represents  $\omega_\beta$  Weights

## 4.7 Discussions

Simulation 1 and Simulation 2 described three major original speech recordings results [76]: late, early and early-late fusion based on scores, features and their combination. Fig. 4.6 is focused on the SIA against all methods used for simulation 1 only, while Fig. 4.7 emphasises the relationship between the mixture size of GMCs and SIA (simulation 1).

In Fig. 4.6 each box plot denotes seven SIA values for different dimension of GMCs  $\{8, 16, 32, 64, 128, 256, 512\}$ . Essentially, the first four boxes (red, pink, green and blue) are used to represent the SIA in the context of multi-bases boxes for different features before using fusion techniques FWMFCC, CMVNMFCC, FWPNCC and CMVNPNC (  $f_1, f_2, g_1, g_2$  ). On the other hand, five boxes show late fusion methods: fusion weights at  $\omega_\beta = 0.9, 0.8$  and  $0.7$  respectively which have been found empirically to yield best SIA as well as maximum and mean fusion; this fusion can be done by fusing the highest SIA features between  $(f_1, f_2)$  which is the  $f_1$  red box with the corresponding highest SIA between  $(g_1, g_2)$  which is the

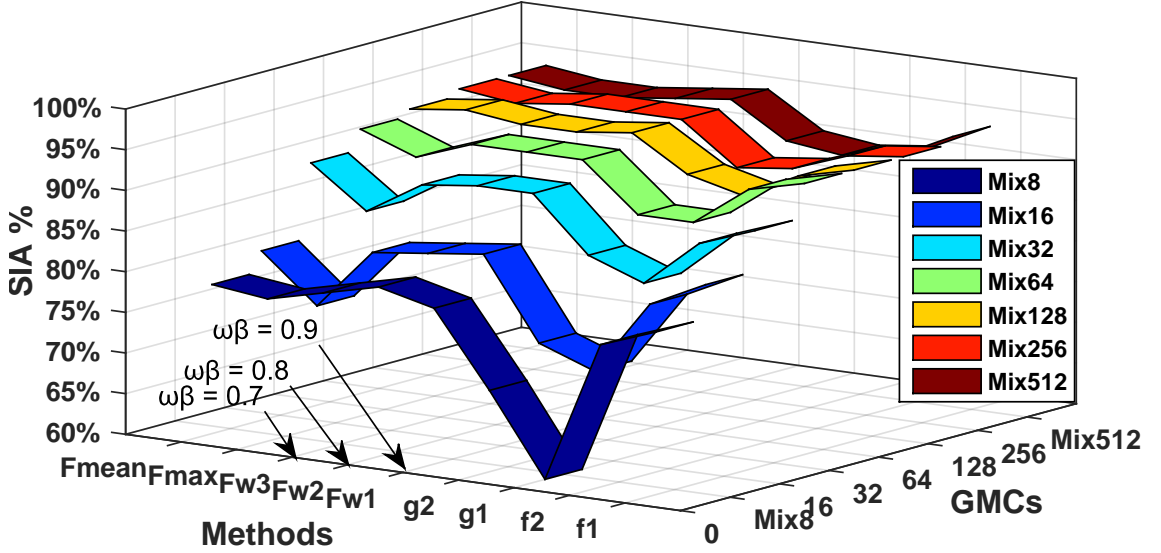


Figure 4.7: Ribbon Plot for Original Speech Recordings Based on Late Fusion Approaches for 16D and Relate to Simulation 1

$g_2$  blue box for all mixture sizes except at mixture size 256 which is represent  $g_1$ . Effectively, fusion boxes have higher SIA against bases boxes, the highest performance is achieved (95%) at weighted sum fusion with  $\omega_\beta = 0.9$  and mixture size 512. It is evident from Fig. 4.6 that the late fusion gives higher SIA than MFCC or PNCC features alone as shown in the multi-bases boxes. However, Fig. 4.7 depicts how far increasing the mixture size will affect the SIA performance as explained in the form of a 3D ribbon plot, it can see when increasing the mixture size from 8 towards mixture size 512 the SIA is growing which is because the benefits are exploited from GMM-UBM and augmented it by fusion methods to improve the SIA. Although, the best SIA result at mixture size 256 is slightly less (94.17%) compared with mixture size 512 (95%) generally the other SIA results for mixture 256 are better than those in mixture 512 [76].

The second original speech recordings simulation results are described in Fig. 4.8 where doubling the feature dimension is accomplished by concatenating different features. In contrast with the first simulation part (simulation 1) 32 feature dimension is used instead of 16. According to Fig. 4.8, there are four bases lines which represent the SIA for the early fusion: BLN1 is concatenated 16 FWMFCC features ( $\mathbf{M}$ ) with the corresponding FWPNCC ( $\mathbf{P}$ ) to produce  $\mathbf{H}_{1,1}$  with 32 features, similarly BLN2 is concatenated 16 CMVNMFCC features ( $\hat{\mathbf{M}}$ ) with the corresponding FWPNCC ( $\mathbf{P}$ ) to produce  $\mathbf{H}_{2,1}$  with 32 features, also in

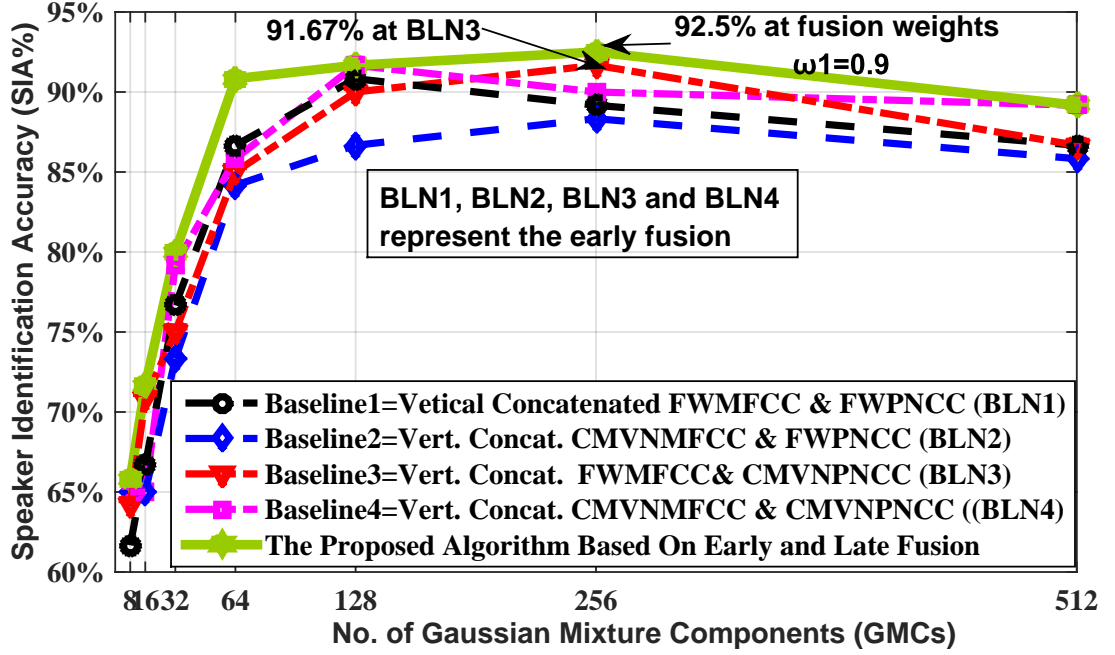


Figure 4.8: Comparison SIA in Original Speech Recordings Based on Early Fusion (Multi-Bases Lines) with Early Late Fusion Relate to simulation 2

BLN3 16 FWMFCC features ( $\mathbf{M}$ ) are concatenated with the corresponding CMVNPNC (  $\hat{\mathbf{P}}$  ) to produce  $\mathbf{H}_{1,2}$  with 32 features and finally in BLN4 the concatenation of 16 CMVNMFCC features ( $\hat{\mathbf{M}}$ ) with the corresponding CMVNPNC ( $\hat{\mathbf{P}}$ ) to produce  $\mathbf{H}_{2,2}$  to yield 32 features as in equation (4.15). However, the second part from simulation 2 is achieved by adding the late fusion methods; the scores which have higher SIA between ( $\mathbf{H}_{2,1}$ ,  $\mathbf{H}_{2,2}$ ) are selected for each mixture size and fused them by late fusion methods with corresponding scores which have the highest SIA between ( $\mathbf{H}_{1,1}$ ,  $\mathbf{H}_{1,2}$ ). Ultimately, the highest SIA from all early-late fusion methods are taken for each mixture to produce the proposed early-late fusion algorithm based on the score-feature combination as explained in Fig. 4.8 as a green line. Furthermore, an early-late fusion algorithm gives slightly higher SIA (92.5%) at mixture size 256 compared with the highest SIA (91.67%) at BLN3 on mixture size 256 as shown in Fig. 4.8 [76].

According to Fig. 4.9, the bar chart demonstrates the comparison between SIA for late fusion (simulation 1) with combination of early-late fusion (simulation 2); empirically the bar chart gives an indication that the late fusion with 16 feature dimension is better (95%) than the early and late fusion which gives higher SIA (92.5%).

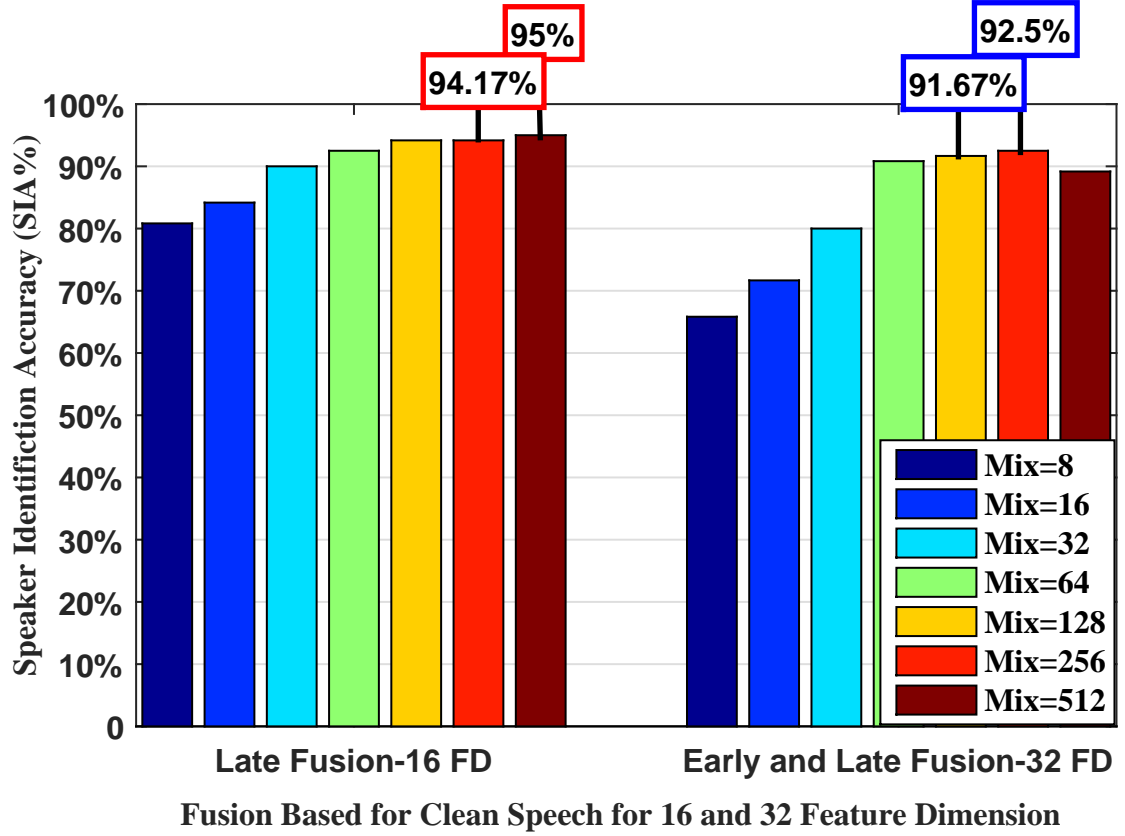


Figure 4.9: Comparison of the Performance for Original Speech Recordings Between Late Fusion (simulation 1) with the Early-Late Fusion (simulation 2)

Simulation 1 and simulation 2 are presented with three techniques of fusion depending on the 16 and 32 Feature Dimensions (FD); the late fusion scheme with (16FD) yielded the best SIA; the early-late with (32FD) the next best and early features fusion (32FD) the lowest SIA [76] [101]. Each fusion method provided higher SIA than the separate normalized MFCC or PNCC features. On the other hand, Fig. 4.10 demonstrates all the highest fusion results are selected for each Gaussian mixture component size were selected from the tables of results to the simulations 1, 2, 3 and 4. It is clear that simulation 1 (Red curve in Fig. 4.10) gives the best performance 95% at mixture size 512, while the second best is achieved by simulation 4 with 93.33% at mixture size 128 (Green curve in Fig. 4.10) followed by simulation 2 with 92.5% at 256 mixture size (Pink curve in Fig. 4.10), whereas the lowest performance accuracy is simulation 3 with 87.5% at both mixture sizes 256 and 512 (Blue curve in Fig. 4.10). Moreover, it is clear from both simulations curves 1 and 4, the SIA curves are gradually increased between mixture size 8 to 64, whilst it seems to be semi-stable for the remaining between mixture size 128 to the 512.

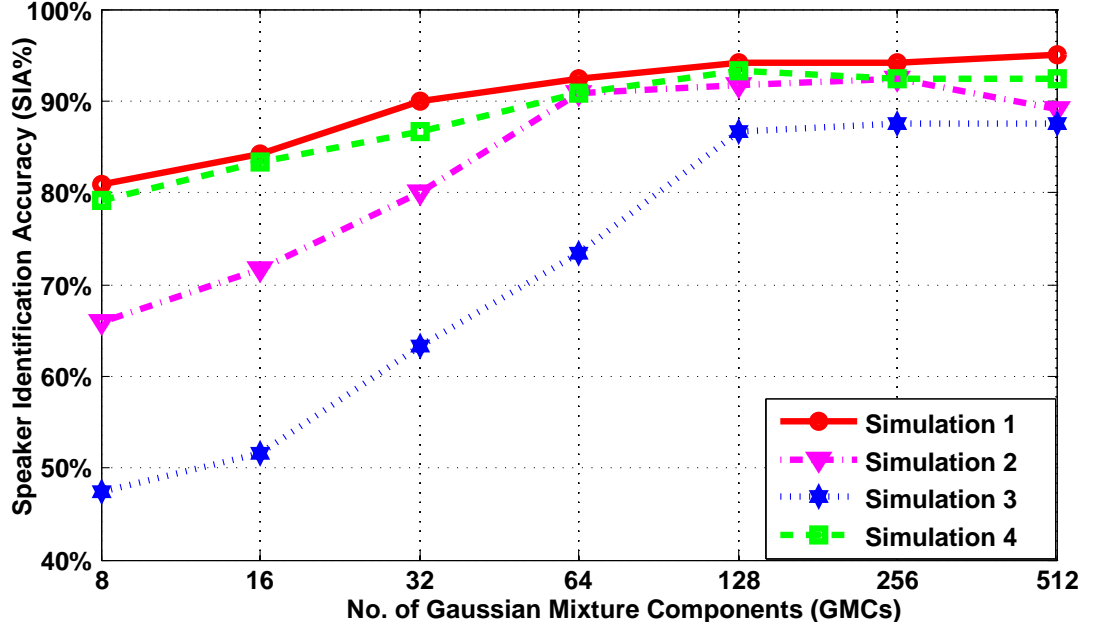


Figure 4.10: SIA Against GMCs for ALL Original Speech Recordings Simulations

However, the SIA for curves in simulations 2 and 3 are dramatically increment for (8-64) mixture sizes, whereas simulation 3 is increasing the SIA to the 128 size then tend to be stable to the end at mixture 512. In contrast, simulation 2 curve is slightly increased for mixtures span 64-256 and tends to reduce the SIA to the end point at mixture size 512. Furthermore, Fig. 4.10 shows empirically the concatenated between the static and dynamic features for the first and second order derivatives in simulation 3 ( the blue dot curve) gives the significant results at mixture sizes (128-512), however the worst SIA can be achieved using this simulation compared with other simulations. In addition, in the noise environment the dynamic features are derivatives in the time domain and represented as a multiplication when converted to the frequency domain and that will increase the noise when the increasing the frequency and this will reduce the performance accuracy for speaker identification. It can be deduced from all simulation tables of results that the mixture size 256 represents the best mixture which give almost the highest SIA.

## 4.8 Summary

In this chapter, four main simulations with fixed original speech recordings length (129250 samples with 8 seconds length) were performed to calculate the speaker identification accuracy for different Gaussian mixture components and different

## 4.8 Summary

---

features dimensions based on different fusion techniques. These fusion methods namely: late fusion (score based 16 FD), early fusion (feature based 32 FD) and early-late fusion (feature-score based 32 FD), concatenated static and dynamic features (feature based 39FD) and finally the multiplication of scores independent for different feature dimensions (16, 32 and 39). This chapter can be summarized by the following points:

- Late fusion is dominant compared with other fusion methods for evaluation the SIA; therefore the late fusion will be considered in the next chapter when applied different environments such as the handset, AWGN and NSN types. In addition, the late fusion consist of three main fusion methods and these are weighted sum, mean and maximum.
- The Gaussian mixture with size 256 gives the highest or second highest SIA for all original speech recordings simulations used in this chapter; therefore the mixture size at 256 is considered for all noisy speech simulations in the next chapter.
- The highest SIA is 95% achieved at mixture size 512 on simulation 1 with late fusion.
- This chapter is used for different feature dimensions such as 16, 32 and 39 and it is clear from simulations the best performance accuracy is accomplished on 16 feature dimension. Subsequently, this feature dimension is used for the all simulations in the next chapter.
- Fusion based for MFCC and PNCC features gives better accuracy than each feature alone.
- Late fusion is the first highest SIA with 95%, while the fusion by multiplication different scores is the second best SIA with 93.33%. Then the combination of early-late is the third order of the best SIA with 92.5% compared with 91.67% achieved by early fusion only and finally the lowest SIA with 87.5% is the late fusion for concatenated static and dynamic features.

## 4.8 Summary

---

Although, the evaluation of this chapter gives improvement related to other work, but the system was still evaluated under ideal acquisition database by using only original speech recordings of TIMIT database. Therefore, various background noise types with and without the handset and their effects on both feature and fusion based will be evaluated in the next chapter with three types of databases: TIMIT, SITW and NIST 2008.

# Chapter 5

## Speaker Identification Using GMM-UBM Approach With Fusion For Challenging Environments With Three Databases

Voice biometrics are used to recognize a person's voice, thereby avoiding the constraints associated with a smart identification card or password recall. In this chapter, a speaker identification system is considered consisting of a feature extraction stage which utilizes both PNCC and MFCC features. Normalization is applied by employing CMVN and FW, together with acoustic modelling using a GMM-UBM. The main contributions are comprehensive evaluations of the effect of both AWGN, and NSN (with and without a G.712 type handset) upon identification performance. In particular, three NSN types with varying SNRs were tested corresponding to: street traffic, a bus interior and a crowded talking environment. The performance evaluation also considered the effect of late fusion techniques based on score fusion, namely mean, maximum, and linear weighted sum fusion. The databases employed were: TIMIT, SITW, and NIST 2008; and 120 speakers were selected from each database to yield 3,600 speech utterances. As recommendations from the study, mean fusion is found to yield overall best performance in terms of SIA with noisy speech, whereas linear weighted sum fusion is overall best for original database recordings.



### 5.1 Background

Speaker identification is one important application of biometrics and forensics to identify speakers based on their unique voice pattern [107], [108] and [109]. According to [37], feature extraction within speaker identification should be less influenced by noise or the person's health. An overview of speaker identification was presented in [3], and increasing the number of speakers and using different types of realistic NSN in evaluation was suggested to develop the field along with exploiting fusion techniques. Despite this research, recognition rate is still a subject of focus. Murty and Yegnanarayana [110] elucidate improvements in a speaker verification system by combining the residual phase derived from linear prediction analysis of the speech signal with the spectral MFCC features. In addition, the NIST 2003 database [110] was used; a 14% EER performance was achieved for MFCC and a 22% rate for the residual phase. Although the combination was better than the individual features alone, the system was not subjected to realistic noise conditions and handset effects. Similar to this approach, Wang et al. [111] used a linear weighted sum for the score fusion but the work did not consider noise, and likewise in [22] channel distortion seems to have been ignored. In [23], different feature combinations were presented using MFCC and LPCC to improve the recognition rate. However, a limited number of speakers was used, only digit speech was employed, and the system was only tested in ideal conditions.

In [58], both the NIST 2008 and TIMIT databases were employed to achieve robust speaker identification and mitigate room reverberation and additive noise, but again handset effects were ignored. Also, to accomplish robust speaker identification, Li and Huang [55] employed CFCCs and used the NTIMIT and Speech Separation Challenge databases, although fusion can also be used to enhance the identification performance. Various neural network based approaches were proposed in [31], without considering different noise and handset conditions. Furthermore, other researchers have employed DNN analysis for speaker identification [49]. In [112], the authors selected 100 speakers from the TIMIT and self-collected databases using Novel Fuzzy Vector Quantization (NFVQ) techniques to enhance the SIS. However, increasing the number of speakers reduced the recognition rate and there was no testing under realistic noise and channel

## 5.1 Background

---

distortion conditions. Moreover, [30] produced a multi-modal neural network by exploiting wavelet analysis, without testing for noise and channel effects and only using 34 speakers. Other researchers have focused on speaker identification and verification applications with background noise to improve and create robust speaker recognition [54]. Khanteymoori et al. [53] utilized a DBN to model speakers and improve identification compared with GMMs, but a limited number of speakers was used. Furthermore, a new discriminative likelihood score weighting technique was proposed for speaker identification, and a likelihood score weighting method was presented for the speaker identification task [113]. In [114], a state of the art speech recognition system was exploited for noisy environments and reverberation. In addition, an empirical study was presented by Reynolds [51], which included the handset variability effects for the speaker recognition purpose using the Switchboard corpus. On the other hand, Reynolds et al. [50] focused on two issues in the speaker identification task, the size of the population and the degradation produced from the noisy telephone channel; their study used the TIMIT and the NTIMIT databases. However, only a limited number of studies have involved a handset, AWGN, and NSN types in conjunction with fusion strategies. In this work, our previous work in [76], [101] was extended with four combinations of features and their score fusion methods for the original recordings; and with AWGN, and three types of NSN: street traffic, bus interior and crowd talk, with and without the G.712 type handset at 16kHz, to provide a wide range of environmental noise conditions. This study emphasise that, although the GMM-UBM approach is well established, no previous study has comprehensively considered three databases, one of which only appeared in 2016, nor the effect of such a wide range of NSN and handset effects.

Section 5.2 contextualises robust biometric speaker identification; Section 5.3 describes adding the noise and applying the handset; Section 5.4 explains the databases and simulation setup; Section 5.5 presents the simulation results and discussions; Section 5.6 includes comparisons with related work; Section 5.7 presents the summary and future work.

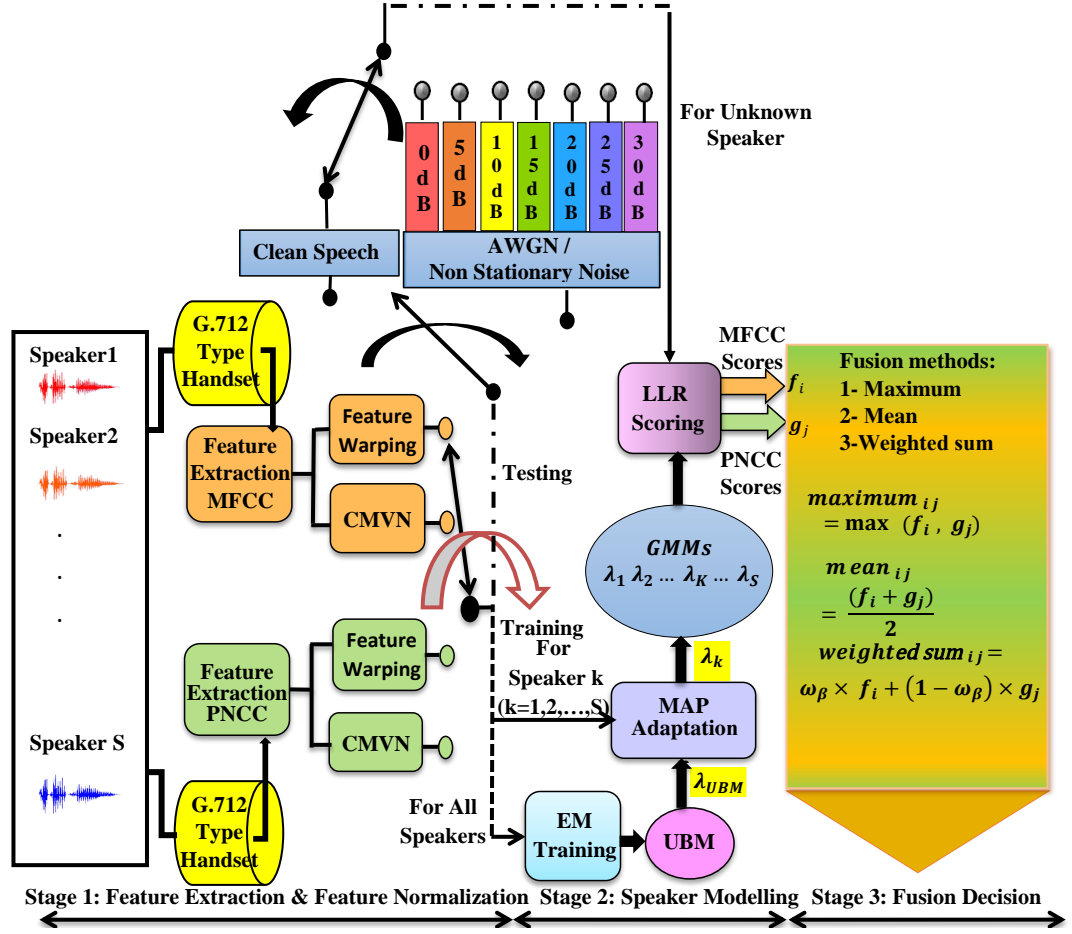


Figure 5.1: Robust Biometric Speaker Identification and Evaluation Framework.

## 5.2 An Overview of a Robust Biometric Speaker Identification System

The main system used in this chapter is represented in Fig. 5.1. The figure has three sections: feature extraction and normalization, speaker modelling and matching, and fusion strategies; it also shows test signals.

### 5.2.1 Feature Extraction and Compensation

In this chapter, to mimic human ear perception, MFCC features are used [115] and combined with the corresponding PNCC features presented in speech recognition systems; these provide robustness [14], and are expected to improve SIA in the presence of background noise. A 16-feature dimension was used to mirror the work in [16] and [76], which used both MFCC and PNCC. In addition, the MFCC features included the zero order  $C_0$  coefficient and the PNCC features, including the  $P_{C_0}$  coefficient. A pre-emphasis FIR filter realising a first order high pass filter was employed to filter the speech samples with emphasis coefficient 0.96 [1]. In addition, framing and Hamming windowing were employed with a frame length of  $16ms$  with an inter-frame overlap of  $8ms$  [13]. Moreover, this work exploits a triangular/MFB and the logarithmic nonlinearity used in MFCC [6], as well as the GFB and power law nonlinearity for PNCC [14], [26] and [27]. This chapter focus on using the PNCC by exploiting the GFB to improve SIA in the presence of stationary AWGN and NSN background noise. In addition, temporal masking, Asymmetric Noise Suppression (ANS) and power law non linearity with a  $1/15$  exponent and GFB were the main elements in the PNCC construction. Further information about PNCC features is provided in [16] [104] [105]. Feature Compensation (Normalization) is widely and effectively used for speaker verification and identification tasks. The main aims of using normalization are to reduce the effects of noise, channel, and handset transducers, and to alleviate linear and nonlinear channel effects. In this study, FW and CMVN over a sliding window are used [18] and [19] to reduce the noise and handset effects, and mitigate linear channel effects; this gives improvements and robustness to SIA [3]. The features and feature normalization are as employed in [76].

### 5.2.2 Speaker Modelling and Matching

#### 5.2.2.1 Gaussian Mixture Model (GMM)

In GMMs, each speaker can be represented by the multivariate parameters of the Gaussian components, namely, mean, covariance and a finite weighted mixture. The weighted sum of the Gaussian mixture components is called a Gaussian mixture density, as presented in equations (4.1) and (4.2) in chapter 4, section 4.2.2 [76]. In this chapter, nodal, diagonal covariance matrices are used instead of full covariance

## 5.2 An Overview of a Robust Biometric Speaker Identification System

---

as used in [3], [76]. In speaker modelling, the EM method estimates parameters for each mixture.

### 5.2.2.2 Gaussian Mixture Model-Universal Background Model (GMM-UBM)

A GMM-UBM was used as in [76] and trained offline with a large amount of data through EM. Furthermore, MAP approach adaptation was employed to train the individual speaker models, and this adaptation was initialized by the UBM and then coupled with the training data for each speaker. The coupling between large training data (UBM) and a small amount of class specific data (individual speaker models) makes the GMM-UBM able to estimate a larger number of parameters which increases the mixture size dimension, and thus the SIA. As in our previous work [101], adaptation coefficients are used in the learning of the means, weights, and variances of the GMM models which can be represented by  $\alpha_i^m, \alpha_i^w, \alpha_i^v$ , where  $i = 1, \dots, S$ , respectively. Furthermore, the adaptation coefficients and parameters used in this chapter are same as presented in chapter 4, section 4.2.2.2.

### 5.2.2.3 Maximum Log-Likelihood Scores

Matching between models built during training and evaluating datasets was carried out by Log Likelihood Ratios (LLRs). In our evaluating studies, 120 speakers were selected from each database. Each speaker has 10 speech utterances, six were employed for training, while the remaining four speech recordings were used for testing. In total, 720 utterances were used for training purpose (6 training files for each of the 120 speakers =  $6 \times 120$ ). In addition, 480 speech utterances were exploited for testing (4 tests for each of the 120 speakers =  $4 \times 120$ ). The model-test set with a length 57,600 represents the multiplication between 120 models with 480 tests ( $120 \times 480$ ). The log-likelihood ratios were calculated as in [76]. Four sets of LLRs were found based on feature and normalization types as described in the next section. A maximum likelihood approach was used to identify speakers as a final decision, as in [3] [42]. The SIA can be calculated as in equation (3.4) [1] [100].

### 5.2.3 Fusion Strategies

Three methods to form a late fusion score were employed as in [76]: weighted sum, maximum and mean fusion. Combined normalization methods were employed to produce normalized MFCC features (FWMFCC and CMVNMFCC). Likewise, normalized methods were used to form PNCC features (FWPNCC and CMVNPNC). Four sets of score vectors could therefore be calculated and are denoted as [101], [76]:  $\mathbf{f}_1$ = Feature Warping MFCC scores vector (FWMFCC),  $\mathbf{f}_2$ = CMVN MFCC scores vector,  $\mathbf{g}_1$ = Feature Warping PNCC scores vector (FWPNCC) and  $\mathbf{g}_2$ = CMVN PNCC scores vector. The maximum, mean, linear weighted sum fusion are defined in chapter 4 in equations (4.15), (4.16) and (4.17), respectively. Further details for fusion strategies can be found in [116] and [117].

## 5.3 Adding Noise and Applying The G.712 Type Handset

### 5.3.1 Adding Stationary AWGN and Non-Stationary Noise

Non Stationary Noise available online from the websites [86] and [85] were used to test the system. Both AWGN and NSN were trimmed to the same fixed length 129,250 speech samples (8 seconds). Different background noise types as well as AWGN were added only in the testing phase with seven SNR levels based on the corresponding noise power (0dB to 30dB) with step size 5dB for each level as in [76].

### 5.3.2 G.712 Type Handset

A G.712 type handset at 16 kHz with a 4<sup>th</sup> order linear IIR filter was derived from the Z transform multiplication of two second order cascaded filters as previously exploited in [3]. The G.712 type handset is applied to the normalized speech signal for both training and testing phases as employed in [76]. The main reason for applying and testing this channel distortion was to achieve robust SIA under original speech recordings, AWGN noisy speech, and realistic NSN conditions. The transfer function of the IIR filter in the z-domain is given as [76]:

### 5.3 Adding Noise and Applying The G.712 Type Handset

$$H(Z) = \frac{b_0 + b_1Z^{-1} + b_2Z^{-2} + b_3Z^{-3} + b_4Z^{-4}}{a_0 + a_1Z^{-1} + a_2Z^{-2} + a_3Z^{-3} + a_4Z^{-4}} \quad (5.1)$$

where the numerator parameters are [1, -0.0216047, -1.92904276, -0.0216047, 1] and denominator parameters are [1, -0.2288945, -1.29745904, 0.06100624, 0.57315888]. The figures for the frequency response and the impulse response for the G.712 handset are added as in Fig. 5.2 and Fig. 5.3. In order to demonstrate the degradation caused by the handset Fig. 5.2 is used and shows there is a degradation in the system bandwidth due to the cut-off frequency and that causes reduction in the speaker identification accuracy.

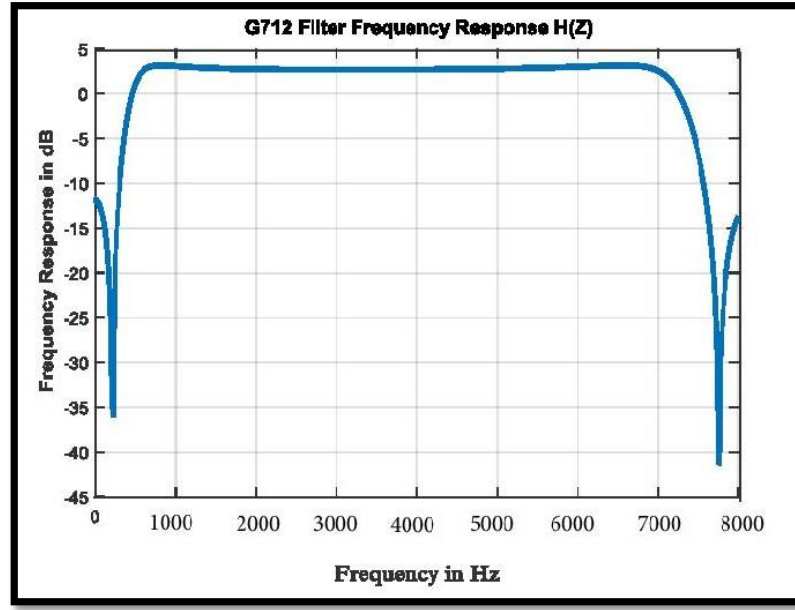


Figure 5.2: Frequency Response for G.712 Type Handset

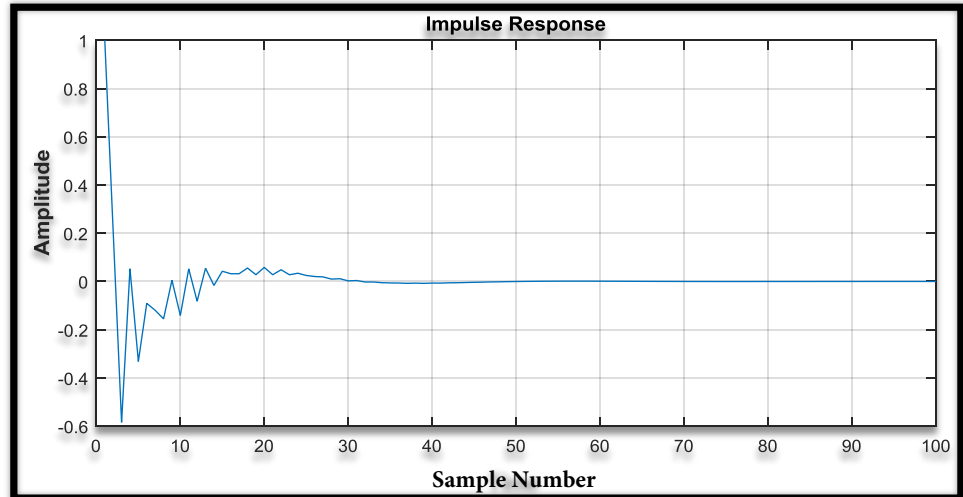


Figure 5.3: Impulse Response for G.712 Type Handset

## 5.4 Databases and Simulation Setups

### 5.4.1 Databases

#### 5.4.1.1 TIMIT Acoustic-Phonetic Continuous Speech Corpus-1993

The TIMIT database is one of the most familiar and widespread speech corpuses used for speech recognition [3], and is available online at the Linguistic Data Consortium website [65]. This corpus has 630 speakers recorded in eight main dialects of American English. In this work, 120 speakers were selected from dialect regions one and four to mirror the work in [1] and the study in [76]. Each speaker has ten speech utterances; six were used for training and four for testing. A fixed speech length of 129,250 samples (8 seconds) was adopted for all 1,200 speech utterances of the 120 speakers, concatenation was used when necessary.

#### 5.4.1.2 The Speakers In The Wild Speaker Recognition Challenge 2016

This challenging database was collected to encourage researchers to develop novel algorithms for benchmarking speaker recognition technology, and is available at [118]. The SITW database was collected under different challenging conditions for open source media: clean interview, outdoor conditions, stadium conditions, and red carpet interviews for single and multi-speakers. In the current study, 120 speakers were selected; most were single speakers, but some were unbalanced multi-speakers. In this case, the target speaker was selected so as to obtain a single speaker, using Goldwave and Audacity software. In addition, each speech file was divided into ten equal lengths, with a fixed length (129,250 samples), to mirror our previous work. However, speech files of less than eight seconds were concatenated to achieve the same fixed length. Six files were used for training and four for testing.

#### 5.4.1.3 2008 NIST Speaker Recognition Evaluation Training Set Part 2-2011

The database is available at [84], and its sources are multilingual telephone and microphone speech of native and bilingual English interview speakers. The sampling frequency was converted from the original 8 kHz to 16 kHz, and 120 English only microphone channel speakers were selected for comparison with the TIMIT and the SITW databases. Again, only single speakers were selected by deleting the



## 5.4 Databases and Simulation Setups

Table 5.1: Parameters and setup used in all experiments and simulations

Aspects	Parameters and experimental setup
Sampling frequency	16000
Window type	Hamming
Frame length	16 ms
Frame shift	8 ms
Pre-emphasis factor	0.96
Databases	<b>TIMIT</b> , <b>SITW</b> and <b>NIST 2008</b>
Number of speakers	120 speakers for each database, total 360 speakers for all databases
Total speech utterances used	1,200 for each database, total 3,600 for all databases
Language	English
Data Source (s)	Microphone Speech for TIMIT and NIST 2008, Hand Annotated Speech from Open Source Media for SITW
No. of samples per speaker	10 for TIMIT, 10 created as well for both SITW and NIST 2008
Testing samples for each database	Total 480 utterances
Training samples for each database	Total 720 utterances
Dialect region	49 speakers are selected from DR1 & 71 speakers are selected from DR4 for TIMIT database to mirror the studies in [76] and [101]
Average sample duration	8 seconds (for each speech utterance in both training and testing); All speech samples were taken with fixed length; concatenation is applied where necessary
Features	MFCC and PNCC
Feature vector dimension	16
Feature normalization	FW and CMVN
Modelling	GMM-UBM
Classifier	LLR
GMCs (Mixtures)	{8, 16, 32, 64, 128, 256, 512 }
Fusion Types	Late Fusion: Mean, Linear Weights, Maximum
System Environment	Original speech recordings, AWGN with G.712 type handset at 16 kHz and (Street-traffic, Bus-interior and Crowd talking NSN) with handset
SNR levels in dB	{0, 5, 10, 15, 20, 25, 30}

interviewers and created six training files and four testing utterances, with a fixed length of eight seconds.

### 5.4.2 Simulation Setups

Six main simulations were performed utilizing the TIMIT, SITW and NIST 2008 databases. Simulation one tested the system without additional noise and handset effects, while simulation two evaluated noisy speech with both AWGN and the G.712 type handset at 16 kHz. Simulations 3-5 employed street traffic, a bus interior and crowd talk NSN, with handset at 16 kHz, respectively. In Simulation 6, Percentage Reduction in SIA (PRSIA) was created to measure the reduction caused by noise and handset effects. Table 5.1 explains the parameters used in the simulations for the three databases, as well as system details, conditions, databases and methods.

## 5.5 Simulation Results and Discussion

In this section, the simulations will be considered in two groups, A and B. Part A includes the five simulations using the three databases: original speech recordings, AWGN with handset, street NSN with handset, bus NSN with handset and crowd talking NSN with handset, respectively. Part B includes further examination of the effects of noise and handset on SIA based on features and fusion methods.

In Part A, Simulation 1 shows the effect of the number of Gaussian Mixture Components (GMCs), namely  $\{8, 16, 32, 64, 128, 256, 512\}$ , upon SIA for speech utterances from the three databases, without noise or a handset. All other simulations in Part A were on noisy speech, with seven SNR levels between (0-30) dB for the same databases at mixture size 256. This noisy speech included the G.712 type handset at 16 kHz under AWGN and three NSN types: street traffic, bus interior and crowd talking.

In Part B, Percentage Reduction in SIA (PRSIA) is used to give further quantitative perspective on each feature type (without fusion) and each fusion technique. In general, all simulations for Part A and Part B present the SIA for the four feature combinations based on MFCC and PNCC, these are: FWMFCC, CMVNMFCC, FWPNCC and CMVNPNC. The scores for the best SIA between the MFCC features (FWMFCC ( $\mathbf{f}_1$ ) and CMVNMFCC ( $\mathbf{f}_2$ )) were fused to obtain the best SIA with the PNCC features (FWPNCC ( $\mathbf{g}_1$ ) and CMVNPNC ( $\mathbf{g}_2$ )).

In Tables 5.2 to 5.6, the row corresponding to fusion decision defines which  $\mathbf{f}$  and  $\mathbf{g}$  vectors yield the highest SIA and therefore only two score vectors were fused. For example, for  $\mathbf{fweight}_{ij}$   $i$  is equal 1 or 2, that means include either  $\mathbf{f}_1$  or  $\mathbf{f}_2$ , and  $j$

is equal 1 or 2 implying using either  $\mathbf{g}_1$  or  $\mathbf{g}_2$ , respectively. For example when the fusion decision is given as  $\mathbf{f}_1$ ,  $\mathbf{g}_1$  and  $\omega_\beta$  equals to 0.9, then  $\mathbf{fweight}_{11} = 0.9 \times \mathbf{f}_1 + 0.1 \times \mathbf{g}_1$ . Their selection is based upon achieving the highest SIA. Furthermore, in this work, mixture sizes of 1024 and 2048 are not considered, because in this work there are insufficient data size for training; utilizing these mixture sizes causes a decline in the SIA performance.

### 5.5.1 Simulations and Experiments for Part A

In all experiments of Part A and Part B, the training and the testing of the GMM-UBM are achieved in total by 120 speakers (1,200 speech utterances are split into 720 for training and 480 for testing) from the TIMIT database in order to produce the SIA for TIMIT. Likewise, the same partitioning method of training and testing, and number of speakers, was applied to both additional databases SITW and NIST 2008.

#### 5.5.1.1 Evaluation of Speech Data from TIMIT, SITW and NIST 2008 Without Handset and Noise (Part A)

In this subsection, Table 5.2 shows the relationship between SIA and GMCs for the three databases according to feature combinations (without fusion), based on MFCC and PNCC features, and various fusion schemes are also considered. According to Table 5.2, the best SIA values were highlighted that achieved using the same fusion decision ( $\mathbf{f}_1$ ,  $\mathbf{g}_2$ ) for all three databases and they are at 95.83% for the mixture size 64, 95% for the mixture size 512 and 82.5% for the mixture size 512 for the NIST 2008, TIMIT and SITW databases, respectively. These best SIAs for the TIMIT and NIST 2008 databases were obtained with weighted sum fusion and  $\omega_\beta$  equal 0.9, while for SITW database the best SIA was also acquired with the weighted sum fusion but with  $\omega_\beta$  equal 0.7. Additionally, from the results of Simulation 1 in Table 5.2, the plots in Fig. 5.4 are formed to give more analysis and discussion. In Fig. 5.4, the highest SIA was selected regardless of using any feature type (without fusion) or fusion method for each mixture size for TIMIT, SITW and NIST 2008 databases. On this basis, the following observations were made. Firstly, increasing the GMCs always increases the SIA for all databases as in the simulations (1 A, 1 B, 1 C), except in mixture size 64 for the NIST 2008

database which obtains better SIA than other mixtures. This is because the GMM-UBM system was trained on a large number of speakers through the UBM, and individual speaker models were adapted through the GMMs. This coupling increases the dimensionality of the GMCs to cover all speakers. Hence, this generally improves the SIA. Secondly, the NIST 2008 evaluation, which is represented by the violet curve in Fig. 5.4 attained the best SIA performance, followed by the red curve for the TIMIT database. In contrast, the evaluation of the SITW database (blue curve) has the lowest SIA performance, as expected, most probably due to the wild and challenging environments compared to the semi-ideal TIMIT database and the less challenging conditions of NIST 2008. Finally, in Fig. 5.4 the NIST 2008 database curve has the smallest variation between the highest SIA (at mixture size 512) and the lowest SIA achieved at mixture size 8. The second smallest variation is for the SITW database. However, the largest variation was attained with the TIMIT database. The main reason for this is that TIMIT is pure clean speech (ideal database as described by [3]), so the highest SIA was achieved with the highest mixture component size (512) which gives very accurate modelling, whereas modelling with the smallest mixture size (8) was not very accurate thereby giving the lowest SIA. On the other hand, for the other databases which do not contain pure speech, such accurate speech modelling is not possible and therefore less variation in SIA as a function of mixture size is generally observed. Firstly, increasing the GMCs mostly increased the SIA for all databases (1 A, 1 B, 1 C), except in mixture size 64 for the NIST 2008 database which gave better SIA comparisons than other mixtures. This may be because the GMM-UBM system was used to train a large number of speakers through the UBM to estimate a larger number of parameters, and individual speaker models were adapted through the GMMs; this coupling increases the dimensionality of the GMCs to cover all speakers, and hence this generally improves the SIA. Secondly, Table 5.2 shows that the NIST 2008 evaluation had the best SIA performance, followed by the TIMIT. In contrast, the evaluation for the SITW database had the lowest SIA performance, as expected, possibly due to the wild and challenging environments compared with the semi ideal TIMIT database and the less challenging conditions of NIST 2008. Finally, the SIA reduction between the lowest GMCs at mixture size 8 and the highest GMCs at mixture size 512 was much lesser with the NIST 2008 than both SITW and TIMIT

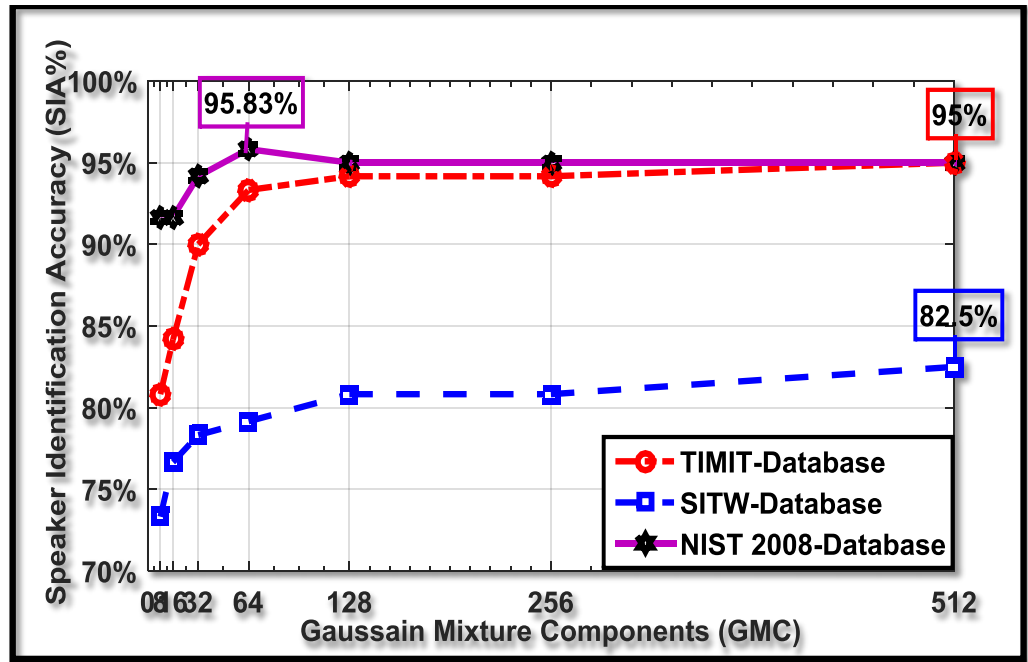


Figure 5.4: Evaluations in Terms of SIA for the TIMIT, SITW and NIST 2008 Databases for Widespread Gaussain Mixture Components {8, 16, 32, 64, 128, 256, 512} Without Handset and Noise Using the GMM-UBM Algorithm

databases.

#### 5.5.1.2 Evaluation of Noisy Speech Data from TIMIT, SITW and NIST 2008 With Handset and Noise (Part A)

This subsection is represented by Tables 5.3 to 5.6, which show the evaluation of TIMIT, SITW and NIST 2008 for noisy speech with handset using different background noises: AWGN, street traffic NSN, bus interior NSN, and crowd talking NSN, respectively. In addition, the handset used in all simulations was the G.712 type handset at 16 kHz. From using time-frequency analysis of the three types of NSN the street traffic and crowd talking have been observed that have broad spectra and therefore have similar effect as AWGN. On the other hand, the dominant energy of the bus-interior noise is low frequency and therefore has least effect on the speech when it is added. Therefore for the AWGN, street and crowd talking, the reduction in SIA performance between 10 and 30 dB was only considered; whereas, for bus-interior, between 0 and 30 dB are considered. According to the tables from Table 5.3 to Table 5.6, the highest SIA results are selected regardless of feature type (without fusion) or fusion method for each SNR level. Then, these results are shown in Fig. 5.5.

## 5.5 Simulation Results and Discussion

Table 5.2: Simulation 1: 1 A, 1 B and 1 C are the SIA for Different Gaussian Mixture Components (GMC) for the TIMIT, SITW and NIST 2008, Respectively

Simulation 1 A: the SIA for Clean Speech <b>TIMIT Database</b>							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
FWMFCC ( $\mathbf{f}_1$ )	80%	84.17%	89.17%	93.33%	93.33%	93.33%	94.17%
CMVNMFCC ( $\mathbf{f}_2$ )	77.5%	80.83%	86.67%	91.67%	91.67%	92.5%	90.83%
FWPNCC ( $\mathbf{g}_1$ )	60%	71.67%	80.83%	86.67%	88.33%	90%	90%
CMVNPNC ( $\mathbf{g}_2$ )	70%	74.17%	83.33%	86.67%	90%	89.17%	90.83%
Fusion Decision	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_1$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )
Fused $\omega_1=0.9$	79.17%	84.17%	90%	91.67%	93.33%	93.33%	<b>95%</b>
Fused $\omega_2=0.8$	80.83%	83.33%	90%	92.5%	93.33%	94.17%	94.17%
Fused $\omega_3=0.77$	80.83%	83.33%	90%	92.5%	93.33%	94.17%	94.17%
Fused $\omega_4=0.7$	79.17%	82.5%	89.17%	92.5%	93.33%	94.17%	93.33%
Fusion Max	77.5%	75%	85%	90%	94.17%	93.33%	93.33%
Fusion Mean	78.33%	80.83%	90%	92.5%	93.33%	94.17%	94.17%

Simulation 1 B: The SIA for <b>SITW Database</b>							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
FWMFCC ( $\mathbf{f}_1$ )	71.67%	75%	76.67%	77.5%	78.33%	78.33%	80%
CMVNMFCC ( $\mathbf{f}_2$ )	69.17%	74.17%	75.83%	78.33%	80.83%	80%	79.17%
FWPNCC ( $\mathbf{g}_1$ )	64.17%	70.83%	78.33%	79.17%	80.83%	79%	79.17%
CMVNPNC ( $\mathbf{g}_2$ )	67.5%	73.33%	77.5%	78.33%	80.83%	80%	80%
Fusion Decision	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_1$ )	( $\mathbf{f}_2, \mathbf{g}_1$ )	( $\mathbf{f}_2, \mathbf{g}_2$ )	( $\mathbf{f}_2, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )
Fused $\omega_1=0.9$	71.67%	75.83%	77.5%	77.5%	80.83%	80.83%	81.67%
Fused $\omega_2=0.8$	71.67%	74.17%	77.5%	77.5%	80.83%	80.83%	81.67%
Fused $\omega_3=0.77$	71.67%	74.17%	76.67%	77.5%	80.83%	80.83%	81.67%
Fused $\omega_4=0.7$	71.67%	75.83%	75.83%	78.33%	80.83%	80.83%	<b>82.5%</b>
Fusion Max	72.5%	75%	77.5%	78.33%	79.17%	78.33%	79.17%
Fusion Mean	73.33%	76.67%	74.17%	79.17%	79.17%	80%	81.67%

Simulation 1 C: The SIA for <b>NIST 2008 Database</b>							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
FWMFCC ( $\mathbf{f}_1$ )	90%	89.17%	92.5%	<b>95.83%</b>	93.33%	92.5%	94.17%
CMVNMFCC ( $\mathbf{f}_2$ )	83.33%	87.5%	88.33%	90.83%	90%	90.83%	89.17%
FWPNCC ( $\mathbf{g}_1$ )	83.33%	86.67%	87.5%	87.5%	89.17%	88.33%	88.33%
CMVNPNC ( $\mathbf{g}_2$ )	84.17%	85%	89.17%	89.17%	89.17%	88.33%	88.33%
Fusion Decision	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_1$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )	( $\mathbf{f}_1, \mathbf{g}_2$ )
Fused $\omega_1=0.9$	89.17%	90.83%	94.17%	<b>95.83%</b>	95%	95%	95%
Fused $\omega_2=0.8$	91.67%	91.67%	93.33%	95%	94.17%	95%	94.17%
Fused $\omega_3=0.77$	90.83%	91.67%	93.33%	94.17%	94.17%	95%	94.17%
Fused $\omega_4=0.7$	90.83%	90.83%	93.33%	94.17%	94.17%	95%	94.17%
Fusion Max	90%	86.67%	93.33%	93.33%	92.5%	92.5%	91.67%
Fusion Mean	88.33%	90%	90.83%	91.67%	92.5%	94.17%	92.5%

## 5.5 Simulation Results and Discussion

Table 5.3: Simulation 2: 2 A, 2 B and 2 C are the SIA Under AWGN and G.712 Type Handset at 16 kHz for Different Signal to Noise Ratio (SNR) Levels for the TIMIT, SITW and NIST 2008, Respectively, at Mixture Size 256

Simulation 2 A: The SIA for Noisy Speech Using <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	0.83%	1.67%	5.83%	14.17%	25.83%	45%	64.17%
CMVNMFCC ( $f_2$ )	0.83%	1.67%	2.5%	5.83%	14.17%	31.67%	57.5%
FWPNCC ( $g_1$ )	1.67%	4.17%	5.83%	15.83%	31.67%	47.5%	60%
CMVNPNC ( $g_2$ )	2.5%	3.33%	7.5%	20%	39.17%	51.67%	60.83%
Fusion Decision	( $f_1, g_2$ )	( $f_1, g_1$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_2$ )
Fused $\omega_1=0.9$	0.83%	1.67%	6.67%	15%	30%	46.67%	66.67%
Fused $\omega_2=0.8$	0.83%	1.67%	5.83%	17.5%	33.33%	45.83%	70%
Fused $\omega_3=0.77$	0.83%	1.67%	5%	17.5%	35%	45.83%	70.83%
Fused $\omega_4=0.7$	0.83%	1.67%	4.17%	16.67%	35.83%	48.33%	70.83%
Fusion Max	2.5%	1.67%	7.5%	16.67%	34.17%	50%	73.33%
Fusion Mean	0.83%	1.67%	6.67%	18.33%	36.67%	51.67%	<b>75.83%</b>

Simulation 2 B: The SIA for Noisy Speech Using <b>SITW Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	3.33%	9.17%	16.67%	31.67%	52.5%	65%	71.67%
CMVNMFCC ( $f_2$ )	3.33%	6.67%	15%	27.5%	47.5%	63.33%	73.33%
FWPNCC ( $g_1$ )	3.33%	6.67%	22.5%	51.67%	71.67%	75.83%	<b>78.33%</b>
CMVNPNC ( $g_2$ )	1.67%	5%	23.33%	53.33%	74.17%	75.83%	<b>78.33%</b>
Fusion Decision	( $f_1, g_1$ )	( $f_1, g_1$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_2, g_2$ )
Fused $\omega_1=0.9$	3.33%	9.17%	18.33%	35.83%	55.83%	71.67%	73.33%
Fused $\omega_2=0.8$	3.33%	10%	20%	38.33%	58.33%	73.33%	75%
Fused $\omega_3=0.77$	3.33%	10%	20%	40.83%	60%	73.33%	75.83%
Fused $\omega_4=0.7$	4.17%	10.83%	21.67%	45%	62.5%	73.33%	76.67%
Fusion Max	4.17%	10%	23.33%	48.33%	62.5%	74.17%	76.67%
Fusion Mean	4.17%	10%	25%	51.67%	73.33%	<b>78.33%</b>	77.5%

Simulation 2 C: The SIA for Noisy Speech Using <b>NIST 2008 Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	0.83%	1.67%	3.33%	7.5%	14.17%	18.33%	20.83%
CMVNMFCC ( $f_2$ )	0.83%	1.67%	2.5%	5%	15.83%	19.17%	23.33%
FWPNCC ( $g_1$ )	0.83%	1.67%	2.5%	2.5%	5.83%	13.33%	25.83%
CMVNPNC ( $g_2$ )	0.83%	1.67%	2.5%	3.33%	5.83%	13.33%	<b>26.67%</b>
Fusion Decision	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_2, g_2$ )	( $f_2, g_2$ )	( $f_2, g_2$ )
Fused $\omega_1=0.9$	0.83%	1.67%	3.33%	7.5%	15.83%	20%	22.5%
Fused $\omega_2=0.8$	0.83%	1.67%	3.33%	6.67%	15.83%	20.83%	23.33%
Fused $\omega_3=0.77$	0.83%	1.67%	3.33%	7.5%	15%	21.67%	24.17%
Fused $\omega_4=0.7$	0.83%	1.67%	3.33%	9.16%	12.5%	21.67%	24.17%
Fusion Max	0.83%	2.5%	3.33%	5%	10.83%	20%	23.33%
Fusion Mean	0.83%	1.67%	3.33%	7.5%	14.38%	18.33%	<b>26.67%</b>

## 5.5 Simulation Results and Discussion

Table 5.4: Simulation 3: 3 A, 3 B and 3 C are the SIA for Street Traffic NSN and G.712 Type Handset at 16 kHz for Different Signal to Noise Ratio (SNR) Levels for the TIMIT, SITW and NIST 2008, Respectively, at Mixture Size 256

<b>Simulation 3 A: The SIA for Street Traffic NSN Using TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	5.83%	15%	26.67%	47.5%	67.5%	78.33%	82.5%
CMVNMFCC ( $f_2$ )	5.83%	15.83%	29.17%	50%	68.33%	79.17%	85%
FWPNCC ( $g_1$ )	1.67%	4.17%	13.33%	30%	40.83%	51.67%	61.67%
CMVNPNC ( $g_2$ )	1.67%	5%	13.33%	35%	50.83%	60%	66.67%
Fusion Decision	( $f_2, g_2$ )	( $f_2, g_2$ )	( $f_2, g_2$ )	( $f_2, g_2$ )	( $f_2, g_2$ )	( $f_2, g_2$ )	( $f_2, g_2$ )
Fused $\omega_1=0.9$	6.67%	18.33%	29.17%	51.67%	72.5%	80.83%	86.67%
Fused $\omega_2=0.8$	5%	18.33%	30.83%	52.5%	73.33%	82.5%	88.33%
Fused $\omega_3=0.77$	5%	17.5%	30%	52.5%	74.17%	82.5%	88.33%
Fused $\omega_4=0.7$	6.67%	17.5%	31.67%	53.33%	73.33%	83.33%	88.33%
Fusion Max	3.33%	9.17%	27.5%	50%	70.83%	82.5%	86.67%
Fusion Mean	2.5%	14.17%	30.83%	55%	73.33%	84.17%	<b>90%</b>

<b>Simulation 3 B: The SIA for Street Traffic NSN Using SITW Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	15.83%	23.33%	41.67%	62.5%	71.67%	76.67%	79.17%
CMVNMFCC ( $f_2$ )	15%	22.5%	32.5%	52.5%	70%	73.33%	75.83%
FWPNCC ( $g_1$ )	5.83%	8.33%	28.33%	50.83%	69.17%	74.17%	78.33%
CMVNPNC ( $g_2$ )	5.83%	9.17%	29.17%	48.33%	69.17%	75.83%	79.17%
Fusion Decision	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_1$ )	( $f_1, g_2$ )	( $f_1, g_2$ )	( $f_1, g_2$ )
Fused $\omega_1=0.9$	15%	24.17%	46.88%	63.33%	70.84%	76.67%	80%
Fused $\omega_2=0.8$	14.17%	24.17%	39.17%	63.33%	71.67%	76.67%	80%
Fused $\omega_3=0.77$	14.17%	24.17%	40%	63.33%	71.67%	76.67%	80%
Fused $\omega_4=0.7$	14.17%	22.5%	39.17%	62.5%	73.33%	77.5%	80%
Fusion Max	10.83%	21.67%	35%	62.5%	70.83%	77.5%	79.17%
Fusion Mean	10.83%	20.83%	35.83%	65%	74.17%	79.17%	<b>81.67%</b>

<b>Simulation 3 C: The SIA for Street Traffic NSN Using NIST 2008 Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	1.67%	2.5%	10.83%	17.5%	29.17%	37.5%	47.5%
CMVNMFCC ( $f_2$ )	1.67%	1.67%	6.67%	12.5%	23.33%	35%	45.83%
FWPNCC ( $g_1$ )	1.67%	2.5%	15%	34.17%	55.83%	74.17%	<b>80%</b>
CMVNPNC ( $g_2$ )	1.67%	1.67%	6.67%	30%	54.17%	71.67%	78.33%
Fusion Decision	( $f_1, g_1$ )	( $f_1, g_1$ )	( $f_1, g_1$ )	( $f_1, g_1$ )	( $f_1, g_1$ )	( $f_1, g_1$ )	( $f_1, g_1$ )
Fused $\omega_1=0.9$	1.67%	5.83%	10.83%	20%	30%	40%	50.83%
Fused $\omega_2=0.8$	1.67%	3.33%	10.83%	21.67%	34.17%	42.5%	55%
Fused $\omega_3=0.77$	1.67%	3.33%	10.83%	22.5%	34.17%	45%	57.5%
Fused $\omega_4=0.7$	1.67%	3.33%	10.83%	24.17%	35.83%	48.33%	60%
Fusion Max	1.67%	3.33%	13.33%	25.83%	39.17%	58.33%	64.17%
Fusion Mean	0.83%	3.33%	12.5%	28.33%	40.83%	50.83%	69.17%



## 5.5 Simulation Results and Discussion

Table 5.5: Simulation 4: 4 A, 4 B and 4 C are the SIA for Bus Interior NSN and G.712 Type Handset at 16 kHz for Different Signal to Noise Ratio Levels for the TIMIT, SITW and NIST 2008, Respectively, at Mixture Size 256

Simulation 4 A: The SIA for Bus Interior NSN Using <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	50.83%	65%	75.83%	79.17%	85%	87.5%	89.17%
CMVNMFCC ( $f_2$ )	53.33%	68.33%	77.5%	82.5%	87.5%	90.83%	<b>91.67%</b>
FWPNCC ( $g_1$ )	10%	23.33%	35.83%	50.83%	65%	70%	72.5%
CMVNPNC ( $g_2$ )	13.33%	27.5%	45%	55.83%	63.33%	69.17%	73.33%
Fusion Decision	$(f_2, g_2)$	$(f_2, g_2)$	$(f_2, g_2)$	$(f_2, g_2)$	$(f_2, g_1)$	$(f_2, g_1)$	$(f_2, g_2)$
Fused $\omega_1=0.9$	55%	69.17%	80.83%	84.17%	88.33%	<b>91.67%</b>	<b>91.67%</b>
Fused $\omega_2=0.8$	56.67%	71.67%	83.33%	85.83%	89.17%	<b>91.67%</b>	90%
Fused $\omega_3=0.77$	56.67%	72.5%	82.5%	85.83%	90%	<b>91.67%</b>	90%
Fused $\omega_4=0.7$	56.67%	70%	83.33%	85.83%	90.83%	90%	89.17%
Fusion Max	40.83%	65%	76.67%	83.33%	84.17%	87.5%	89.17%
Fusion Mean	51.67%	68.33%	78.33%	84.17%	86.67%	88.33%	90.83%

Simulation 4 B: The SIA for Bus Interior NSN Using <b>SITW Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	65.83%	70.83%	73.33%	75.83%	77.5%	79.17%	79.17%
CMVNMFCC ( $f_2$ )	66.67%	70.83%	72.5%	73.33%	76.67%	77.5%	79.17%
FWPNCC ( $g_1$ )	27.5%	49.17%	64.17%	71.67%	75.83%	77.5%	79.17%
CMVNPNC ( $g_2$ )	28.33%	48.33%	65%	72.5%	75%	79.17%	80%
Fusion Decision	$(f_2, g_2)$	$(f_2, g_1)$	$(f_1, g_2)$	$(f_1, g_2)$	$(f_1, g_1)$	$(f_1, g_2)$	$(f_1, g_2)$
Fused $\omega_1=0.9$	66.67%	71.67%	73.33%	75.83%	77.5%	80%	80%
Fused $\omega_2=0.8$	65%	72.5%	74.17%	75.83%	77.5%	80.83%	80%
Fused $\omega_3=0.77$	66.67%	72.5%	75%	76.67%	77.5%	<b>81.67%</b>	80%
Fused $\omega_4=0.7$	65.83%	72.5%	75%	76.67%	78.33%	80.83%	80.83%
Fusion Max	63.33%	72.5%	73.33%	79.17%	80%	80.83%	80.83%
Fusion Mean	59.17%	70.83%	73.33%	76.67%	79.17%	<b>81.67%</b>	80.83%

Simulation 4 C: The SIA for Bus Interior NSN Using <b>NIST 2008 Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	22.5%	32.5%	36.67%	42.5%	59.17%	72.5%	85.83%
CMVNMFCC ( $f_2$ )	19.17%	28.33%	36.67%	45%	60%	74.17%	85.83%
FWPNCC ( $g_1$ )	7.5%	15%	37.5%	57.5%	71.67%	80%	80%
CMVNPNC ( $g_2$ )	6.67%	14.17%	35.83%	57.5%	73.33%	82.5%	84.17%
Fusion Decision	$(f_1, g_1)$	$(f_1, g_1)$	$(f_2, g_1)$	$(f_2, g_2)$	$(f_2, g_2)$	$(f_2, g_2)$	$(f_2, g_2)$
Fused $\omega_1=0.9$	20.83%	32.5%	39.17%	49.17%	60.83%	78.33%	88.33%
Fused $\omega_2=0.8$	17.5%	30%	40.83%	53.33%	63.33%	84.17%	90%
Fused $\omega_3=0.77$	17.5%	27.5%	40.83%	53.33%	64.17%	84.17%	90.83%
Fused $\omega_4=0.7$	17.5%	26.67%	42.5%	54.17%	68.33%	83.33%	90.83%
Fusion Max	15%	28.33%	41.67%	53.33%	69.17%	85%	89.17%
Fusion Mean	15.83%	25.83%	45.83%	58.33%	75%	86.67%	<b>92.5%</b>

## 5.5 Simulation Results and Discussion

Table 5.6: Simulation 5: 5 A, 5 B and 5 C are The SIA for Crowded Talking NSN and G.712 Type Handset at 16 kHz for Different Signal to Noise Ratio Levels for the TIMIT, SITW and NIST 2008, Respectively, at Mixture Size 256

<b>Simulation 5 A: The SIA for Crowded Talking NSN Using TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	9.17%	18.33%	35%	50.83%	66.67%	74.17%	80%
CMVNMFCC ( $f_2$ )	7.5%	19.17%	34.17%	55.83%	69.17%	81.67%	87.5%
FWPNCC ( $g_1$ )	1.67%	2.5%	15.83%	29.17%	43.33%	56.67%	59.17%
CMVNPNC ( $g_2$ )	1.67%	5%	19.17%	35%	54.17%	60.83%	68.33%
Fusion Decision	$(f_1, g_2)$	$(f_2, g_2)$	$(f_1, g_2)$	$(f_2, g_2)$	$(f_2, g_2)$	$(f_2, g_2)$	$(f_2, g_2)$
Fused $\omega_1=0.9$	10%	19.17%	35.83%	57.5%	70.83%	83.33%	87.5%
Fused $\omega_2=0.8$	10%	16.67%	36.67%	59.17%	71.67%	83.33%	90%
Fused $\omega_3=0.77$	10%	16.67%	36.67%	60%	72.5%	83.33%	88.33%
Fused $\omega_4=0.7$	8.33%	16.67%	37.5%	61.67%	74.17%	84.17%	88.33%
Fusion Max	2.5%	9.17%	39.17%	52.5%	73.33%	84.17%	88.33%
Fusion Mean	5%	15%	38.33%	62.5%	73.33%	82.5%	89.17%

<b>Simulation 5 B: The SIA for Crowded Talking NSN Using SITW Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	18.33%	33.33%	45.83%	64.17%	73.33%	75.83%	78.33%
CMVNMFCC ( $f_2$ )	15.83%	30%	43.33%	59.17%	72.5%	75.83%	77.5%
FWPNCC ( $g_1$ )	5%	15%	33.33%	59.17%	71.67%	76.67%	79.17%
CMVNPNC ( $g_2$ )	4.17%	12.5%	30%	53.33%	70%	75.83%	80.83%
Fusion Decision	$(f_1, g_1)$	$(f_1, g_1)$	$(f_1, g_1)$	$(f_1, g_1)$	$(f_1, g_1)$	$(f_1, g_1)$	$(f_1, g_2)$
Fused $\omega_1=0.9$	20%	65%	48.33%	67.5%	73.33%	75.83%	80%
Fused $\omega_2=0.8$	18.33%	61.67%	50%	68.33%	73.33%	75.83%	80%
Fused $\omega_3=0.77$	17.5%	60%	50.83%	69.17%	73.33%	75.83%	80%
Fused $\omega_4=0.7$	17.5%	57.5%	53.33%	70%	73.33%	77.5%	80%
Fusion Max	14.17%	48.33%	46.67%	65.83%	73.33%	76.67%	80.83%
Fusion Mean	11.67%	45%	50.83%	72.5%	75%	78.33%	82.5%

<b>Simulation 5 C: The SIA for Crowded Talking NSN Using NIST 2008 Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
FWMFCC ( $f_1$ )	7.5%	12.5%	24.17%	30%	37.5%	47.5%	66.67%
CMVNMFCC ( $f_2$ )	3.33%	10.83%	18.33%	28.33%	40.83%	46.67%	67.5%
FWPNCC ( $g_1$ )	3.33%	11.67%	29.17%	29.17%	67.5%	78.33%	80.83%
CMVNPNC ( $g_2$ )	2.5%	10%	24.17%	24.17%	68.33%	79.17%	82.5%
Fusion Decision	$(f_1, g_1)$	$(f_1, g_1)$	$(f_1, g_1)$	$(f_1, g_2)$	$(f_2, g_2)$	$(f_1, g_2)$	$(f_2, g_2)$
Fused $\omega_1=0.9$	6.67%	15%	24.17%	34.17%	45.83%	55.83%	70.83%
Fused $\omega_2=0.8$	10%	15%	24.17%	35%	48.33%	60.83%	75.83%
Fused $\omega_3=0.77$	10%	15%	25.83%	36.67%	49.17%	61.67%	77.5%
Fused $\omega_4=0.7$	10%	15%	28.33%	40.83%	49.17%	64.17%	80%
Fusion Max	8.33%	15.83%	29.17%	45.83%	51.67%	70%	77.5%
Fusion Mean	8.33%	17.5%	30%	45%	57.5%	73.33%	84.17%

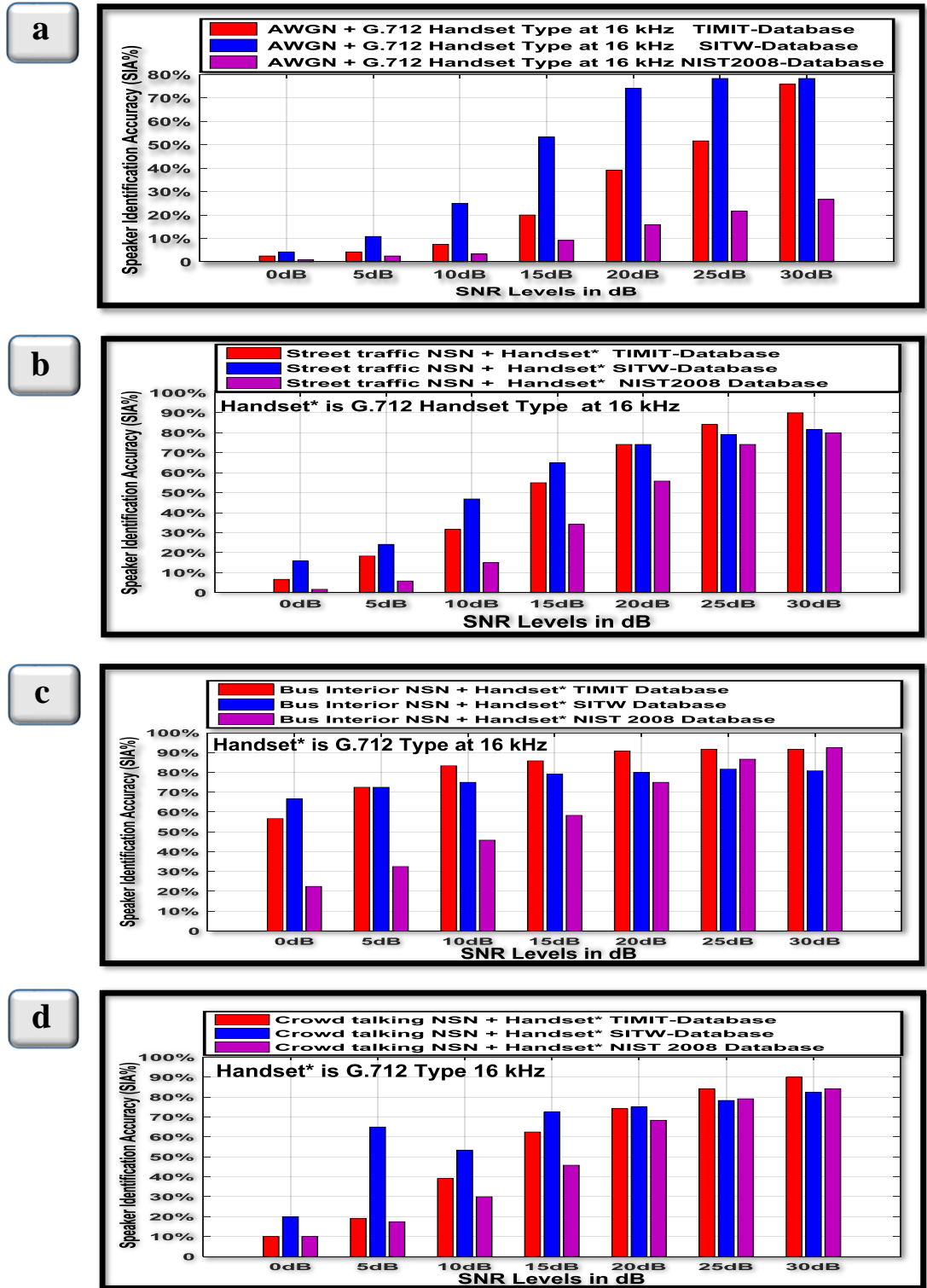


Figure 5.5: Performance Measurement for Noisy Speech for the TIMIT, SITW and NIST 2008 Database at Mixture Size 256 Under G.712 Type Handset at 16 kHz With Background Noise (a) AWGN , (b) Street Traffic NSN, (c) Bus-Interior and (d) Crowd Talking NSN for Wide Range of SNR Levels (0-30) dB and Using GMM-UBM Algorithm.

## 5.5 Simulation Results and Discussion

---

Firstly, for AWGN and G.712 type handset, represented in Table 5.3, the bar charts in Fig. 5.5 (a) can be used to analyse and discuss the results given in Table 5.3. The figure shows the reduction in SIA was from 75.83% at 30 dB to 7.5% at 10 dB for the TIMIT database, while in SITW the reduction in the SIA was from 78.33% at 30 dB to 25% at 10 dB. In contrast, the NIST 2008 had the lowest SIA among all other databases at 30 dB with 26.67% then this was reduced to the 3.33% at 10 dB, as such all databases were affected by stationary noise, with a constant spectrum profile. The particular sensitivity to such noise when applied to the NIST 2008 database may be due to the natural characteristics of the interview speech.

Secondly, for street traffic NSN with handset, seen in Table 5.4, Fig. 5.5 (b) shows that the reduction in SIA was from 90% at 30 dB to 31.67% at 10 dB for the TIMIT database. Similarly, the reduction in SIA obtained by the NIST 2008 database was from 80% to 15%. In contrast, the lowest reduction in the performance accuracy was attained using the SITW database with SIA 81.67% at 30 dB dropping down to 46.88% at 10 dB. As a consequence, the SITW database has the lowest reduction in SIA compared with the other three databases used for the evaluation.

Thirdly, for the bus interior NSN, seen in Table 5.5, Fig. 5.5 (c) illustrates that the reduction in SIA was from 91.67% at 30 dB to 56.67% at 0 dB for the TIMIT database. Likewise, for the SITW database the SIA reduction was from 80.83% to 66.67% for 30 dB and 0 dB, respectively. However, the highest reduction in SIA was for the NIST 2008 database with SIA 92.5% at 30 dB to 22.5% at 0 dB.

Finally, the results in Table 5.6, Fig. 5.5 (d) show that the evaluation of the crowd talking NSN with the handset evaluation was similar to the street NSN. For the TIMIT database, the reduction in SIA was from 90% at 30 dB to 39.17% at 10 dB. Similar to this reduction the figure for the NIST 2008 database were 84.17% at 30 dB to 30% at 10 dB. In contrast, for the SITW database the reduction in SIA was from 82.5% to 53.33%. Considering the reduction in SIA for all simulations as a result of noise and handset effects, the most important issue is the relative sensitivities of the various methods to the environments. To address this point, further comparative analysis are considered.

### 5.5.2 Simulations and Experiments for Part B

In this study, based on the feature types (using four feature combinations without fusion) and fusion methods, the quantitative perspectives were measured by calculating the PRSIA.

#### 5.5.2.1 Quantitative Perspective for Noise and Handset Effects in PartB

The PRSIA was calculated for different conditions as in equation (5.2):

$$PRSIA_{cond} = \frac{SIA_{cond} - SIA_{clean}}{SIA_{clean}} \times 100\% \quad (5.2)$$

where:  $cond \in \{1, 2, 3, 4\}$ , 1 refers to the AWGN and handset, 2 to street traffic NSN and handset, 3 to the bus interior NSN and handset, and 4 to the crowded talking NSN and handset. The handset used was G.712 type at 16 kHz. This equation measured the  $SIA_{clean}$  at mixture size 256 for the original recordings in TIMIT, SITW and NIST 2008, without noise and handset conditions. Then the  $SIA_{cond}$  was measured under the four conditions in the testing phase. Table 5.7 presents the results of PRSIA for each condition, depending on the noise type with handset, each feature type, and each fusion method. The negative sign “-” refers to reduction, while “+” refers to increase. It is surprising to see a few positive sign values in Table 5.7, as different background noise with handset effects are considering, and the system should generally be degraded; but at SNR 30 dB the very small amount of noise may have a stabilization effect on the speaker identification system. Moreover, all positive sign values in Table 5.7 are for the challenging new database (SITW). Generally, however, from Table 5.7 all the results for TIMIT and NIST 2008 can noticed at SNR 30 dB have negative sign values, meaning a reduction in the SIA as a result of the noise and handset effects. Secondly, most of the fusion methods reduced the PRSIA for all databases used.

Further, and most importantly, NIST 2008 is more sensitive to noise, especially AWGN, and has a higher reduction in PRSIA compared with TIMIT and SITW. In contrast, SITW seems relatively robust against noise. The fusion mean seems to have the lowest reduction in SIA compared with other fusion methods. However, MFCC features have less reduction in SIA for the TIMIT database, while this position is reversed for SITW and NIST 2008. For PNCC, the features have less reduction than MFCC in terms of the SIA. Finally, the highest reduction in all databases occurred under the AWGN with handset condition, which is due to the uniformity of the spectrum effect of the noise. The bus interior NSN and handset has the lowest reduction which as stated earlier is due to its low frequency nature. The results for other noise conditions (street and crowded talking) are between the AWGN and bus NSN effects.

## 5.6 Related Works Based on the Proposed Speaker Identification System

Table 5.8 summarizes results mostly at SNR 30 dB, where Cond.1 is speech files from TIMIT, SITW and NIST 2008 without handset and noise, termed original speech recordings; Cond.2 is noisy speech by AWGN and handset; Cond.3 is street NSN and handset; Cond.4 is bus NSN and handset; Cond.5 is crowded talking NSN and handset. The handset used in all noise conditions is G.712 type at 16 kHz. Comparisons show improvement in SIA with the TIMIT database in cond.1 over the state of the art methods due to Kumar et al. [1] and Togneri and Pullella [3]. However, Ming et al. in their earlier work in [52] attain higher SIA in Cond.1 with TIMIT but only with a GMM model and 630 speakers, but they do not consider a handset in Cond. 3. New benchmark figures contributed from this study for a range of environmental noise conditions with the three databases are provided by Cond.2 - Cond.5.

## 5.7 Summary

Table 5.7: Percentage Reduction in SIA (PRSIA) for the TIMIT, SITW and NIST 2008, Respectively, Under G.712 Type Handset at 16 kHz at Mixture Size 256 and SNR 30 dB, AWGN, Street Traffic, Bus Interior, Crowded Talking NSN

Simulation 6 A: PRSIA for <b>TIMIT Database</b>					
Environments	Methods	AWGN-WH	Street-WH	Bus-WH	Crowd -WH
Feature based	FWMFCC ( $f_1$ )	-31.24%	-11.6%	-4.46%	-14.28%
	CMVNMFCC ( $f_2$ )	-37.84%	-8.11%	-0.9%	-5.41%
	FWPNCC ( $g_1$ )	-33.33%	-31.48%	-19.44%	-34.26%
	CMVNPNC ( $g_2$ )	-31.78%	-25.23%	-17.76%	-23.37%
Fusion based	Fused $\omega_1=0.9$	-28.57%	-7.14%	-3.57%	-6.25%
	Fused $\omega_2=0.8$	-25.67%	-6.2%	-4.43%	-4.43%
	Fused $\omega_3=0.77$	-24.78%	-6.2%	-4.43%	-6.2%
	Fused $\omega_4=0.7$	-24.78%	-6.2%	-5.31%	-6.2%
	Fusion Max	-21.43%	-7.14%	-4.46%	-5.36%
	Fusion Mean	-19.48%	-4.43%	-3.55%	-5.31%

Simulation 6 B: PRSIA for <b>SITW database</b>					
Environments	Methods	AWGN-WH	Street-WH	Bus-WH	Crowd -WH
Feature based	FWMFCC ( $f_1$ )	-8.5%	+1.07%	+1.07%	0%
	CMVNMFCC ( $f_2$ )	-8.34%	-5.21%	-3.13%	-3.13%
	FWPNCC ( $g_1$ )	-0.85%	-0.85%	-1.9%	+0.22%
	CMVNPNC ( $g_2$ )	-2.09%	-1.04%	-1.04%	+1.04%
Fusion based	Fused $\omega_1=0.9$	-9.28%	-1.03%	-1.02%	-1.03%
	Fused $\omega_2=0.8$	-7.21%	-1.03%	-1.02%	-1.03%
	Fused $\omega_3=0.77$	-6.19%	-1.03%	-1.02%	-1.03%
	Fused $\omega_4=0.7$	-5.15%	-1.03%	0%	-1.03%
	Fusion Max	-2.12%	+1.07%	+3.19%	+3.19%
	Fusion Mean	-3.13%	+2.09%	+2.09%	+3.13%

Simulation 6 C: PRSIA for <b>NIST 2008 database</b>					
Environments	Methods	AWGN-WH	Street-WH	Bus-WH	Crowd -WH
Feature based	FWMFCC ( $f_1$ )	-77.48%	-48.65%	-7.21%	-27.92%
	CMVNMFCC ( $f_2$ )	-74.31%	-49.54%	-5.5%	-25.69%
	FWPNCC ( $g_1$ )	-70.76%	-9.43%	-9.43%	-8.49%
	CMVNPNC ( $g_2$ )	-69.81%	-11.32%	-4.71%	-6.6%
Fusion based	Fused $\omega_1=0.9$	-76.32%	-46.49%	-7.02%	-25.44%
	Fused $\omega_2=0.8$	-75.44%	-42.11%	-5.26%	-20.18%
	Fused $\omega_3=0.77$	-74.56%	-39.47%	-4.39%	-18.42%
	Fused $\omega_4=0.7$	-74.56%	-36.84%	-4.39%	-15.79%
	Fusion Max	-78.38%	-30.63%	-3.6%	-16.22%
	Fusion Mean	-71.68%	-26.55%	-1.77%	-10.62%

## 5.7 Summary

In this study, a comprehensive evaluation was provided of text independent closed set speaker identification in the presence of AWGN and NSN types with a G.712 type handset at 16 kHz, to provide benchmark evaluations of three different databases. Different feature combinations are presented based on MFCC and PNCC, modelled by the GMM-UBM approach with and without fusion techniques

(maximum, mean and weighted sum fusion). The evaluations were conducted under challenging environments including in the presence of the G.712 handset, AWGN, and various NSN types. Three databases (TIMIT, NIST 2008 and SITW) with a wide range of seven SNR levels (0-30) dB with step size 5 dB were employed. In addition, a wide range of Gaussian mixture components {8, 16, 32, 64, 128, 256, 512} for original speech recordings was also considered. Thorough evaluation and results were provided by this research in order to give benchmark evaluations and results for the three databases for other researchers working in the speaker identification area. The major findings from this study are:

- *On the basis of the evaluations of three databases without the noise and handset conditions, the best speaker identification method for all three databases used was weighted sum fusion.*
- *Based on the three databases without the noise and handset conditions, the order for best SIAs were: NIST2008, TIMIT, SITW with 95.83%, 95% and 82.5%, respectively at mixture sizes 64, 512 and also 512, respectively. These SIAs were achieved by using weighted sum fusion with 90 percent from FWMFCC features and 10 percent from the corresponding CMVNPNC features for both the TIMIT and NIST 2008 database. On the other hand, in the SITW database 70 percent from FWMFCC features was fused with 30 percent from the corresponding CMVNPNC features. The weighting should therefore be chosen as a function of the fidelity of the speech recordings.*
- *On the basis of the results in this chapter, the evaluations in noisy conditions suggest that mean fusion of four combinations of two types of features from (FWMFCC, CMVNMFCC, FWPNC and CMVNPNC) is the most robust method for a practical speaker identification system, but there is not a consistent best pairing.*

The next chapter will consider a similar extensive evaluation for a speaker identification system built from an I-vector approach [37].



Table 5.8: Comparisons with the State of the Art of SIA

Authors	Database	System Approach	Cond.1	Cond.2	Cond.3	Cond.4	Cond.5
<b>Proposed</b> In this chapter	<b>TIMIT</b> 120 speakers Microphone channel	Fusion Based GMM-UBM	<b>95%</b>	<b>75.83%</b> SNR 30dB	<b>90%</b> SNR 30dB	<b>91.67%</b> SNR 30dB	<b>90%</b> SNR 30dB
<b>Proposed</b> In this chapter	<b>SITW</b> 120 speakers	Fusion Based GMM-UBM	<b>82.5%</b>	<b>78.33%</b> SNR 30dB	<b>81.67%</b> SNR 30dB	<b>81.67%</b> SNR 30dB	<b>82.5%</b> SNR 30dB
<b>Proposed</b> In this chapter	<b>NIST 2008</b> 120 speakers Microphone channel	Fusion Based GMM-UBM	<b>95.83%</b>	<b>26.67%</b> SNR 30dB	<b>80%</b> SNR 30dB	<b>92.5%</b> SNR 30dB	<b>84.17%</b> SNR 30dB
Kumar et al. [1] [2012]	TIMIT 120 speakers	GMM	93.88%				
Togneri and Pullella [3] [2011]	TIMIT 64 speakers	GMM-UBM	94.5%	74.2% at SNR 30dB			
Ming et al. [52] [2007]	TIMIT 630 speakers	GMM Mix 128	96.51%		92.86% at 20 dB without handset		

## Chapter 6

# Fusion-based Speaker Identification Using Multi-Dimensional I-vectors in Challenging Environments for Four Databases

I-vectors represent the state of the art, especially for speaker recognition applications, and yet few researchers have exploited fusion-based I-vectors for this task. In this chapter, a novel fusion I-vectors with classification approach using an ELM is employed. The system is tested using original speech recordings and various types of NSN as background context, including street traffic NSN, a bus-interior NSN, and crowd talking NSN, as well as AWGN. In addition, the evaluation includes G.712 type handset effects at 16 kHz. Hence, four I-vector combinations are achieved using CMVN and FW to the MFCC, and PNCC features. Various fusion I-vectors are employed to improve the identification accuracy, and the ELM is exploited to identify speakers; in this way, SIA is calculated. The system is evaluated with four different databases: the 2016 SITW database, the NIST 2008, the TIMIT and the NTIMIT. From each database, 120 speakers with 1,200 speech utterances are used (overall 480 speakers with 4,800 speech utterances are used in this chapter). A limited experiment is also performed in this chapter using the original speech recordings from NTIMIT

database without noise conditions or handset effects. The proposed system is compared with the GMM-UBM and other state of the art approaches. The results show that the I-vector approach outperforms the GMM-UBM approach and other state of the art methods under specific conditions. This chapter also gives fair comparisons in terms of the SIA for the different databases with a wide range of UBM mixture sizes and seven SNR levels for AWGN and different NSN types. No previous study has comprehensively considered four databases, nor the effect of such a wide range of NSN and handset effects.

## 6.1 Background

The I-vector is a recent and most interesting state of the art method initially proposed by Dehak et al. in 2011 [119] for speaker verification applications. Prior to the I-vector, the traditional approaches for modelling speakers were JFA [34], GMM-UBM, supervector GMM, and GMM. Some researchers were interested in text independent speaker recognition such as in [120] and [121], while others focused on speaker dependent recognition [122]. However, the majority of researchers have concentrated on exploiting the modern modelling approach using I-vectors for speaker verification [16, 35, 36, 123–128]. They have also proposed different compensation methods for channel variabilities, such as WCCN, NAP, and LDA. For system classification, PLDA, SRC, CDS and SVM have been proposed. The NIST 2008 database plays an important role in the evaluation of the state of the art methods, especially the I-vector approach with different dimensions (400, 800, 1200 and 1600). In addition, Deep Belief Networks (DBNs) were considered to improve the I-vector system in [129] and [130], respectively. Further research on I-vector speaker recognition and verification applications were made in [131–135], and in [136] for spoof voice verification. JFA is another approach in the presence of noise, and fusion based speaker verification, as in [137]. More generally, a survey of speech processing, including the I-vector approach to the concept, applications, challenges, approaches and hybrid approaches, as well as the main toolkits, were made in [37].

However, various researchers have exploited I-vectors for other applications; [138] and [139] proposed emotional speaker and speech recognition based on I-vectors; the interesting point in [138] is the comparison of the I-vector results with

## 6.1 Background

---

GMM-UBM. Another application for I-vectors was suggested in [140], which sought to estimate the age of speakers. In [141], an I-vector framework was utilized within a MAP approach estimation with additive noise to achieve robustness in speaker recognition using the NIST 2008 database. Previous studies have also focused on using a fusion based GMM-UBM for speaker identification with different feature combinations basically constructed from MFCC and PNCC, such as [76, 101] which both evaluated 120 speakers from the TIMIT database for original speech recordings and AWGN with and without a handset, including feature and score fusion.

The main contributions of this chapter are as follows. Firstly, a new text independent closed set speaker identification system is established which mainly consists of four feature combinations of I-vectors with 100 and 200 I-vector dimensions without fusion, consisting of the normalized MFCC and PNCC features (FW and CMVN), mainly FWMFCC, CMVNMFCC, FWPNCC and CMVNPNCC. Then, I-vector fusion with the two normalized MFCC features with the corresponding normalized PNCC features were applied using maximum, mean, weighted sum and cumulative fusion and have the same dimension for the feature combinations of the I-vector (without fusion), to improve the SIA. Similarly, both interleaved and concatenated I-vector fusion were used with doubled I-vector dimensions, as a consequence of fusion, to produce 200 and 400 I-vector dimensions. On the other hand, fusion of the four combinations of I-vectors together was achieved (FWMFCC, CMVNMFCC, FWPNCC and CMVNPNCC), using concatenated fusion to yield 400 and 800 I-vector dimensions. Furthermore, interleaved and concatenated fusion I-vectors were employed for the different hidden layer nodes with the ELM, to classify the genuine speakers with the fusion-based I-vectors. Secondly, the proposed system is evaluated with four different databases using 120 speakers (1,200 speech utterances) from each database (equalling 480 speakers with 4,800 speech utterances from all the experiments). The novel 2016 SITW challenge database, and the 2008 NIST database, were exploited for speaker identification. Moreover, American English speakers from the popular and widely available TIMIT database [65], which is also called TIMIT Acoustic-Phonetic Continuous Speech Corpus was also considered [3]. Also, a further small experiment was undertaken using a telephone bandwidth version called NTIMIT. Thirdly, the system was evaluated under

## 6.1 Background

---

original speech recordings, AWGN, and various types of NSN namely, street-traffic NSN, bus-interior NSN and crowd talking NSN; and all noise was added in the testing phase. However, the G.712 type handset at 16 kHz was applied to both phases (training and testing). This research uses a wide range of SNR levels and UBM mixture sizes. Moreover, the system provides fair comparisons with the corresponding fusion-based GMM-UBM system and other state of the art methods. In this work, the I-vector is exploited for a lower dimension compared with standard research, because the training and testing is achieved for each database alone. Therefore there was no need to compensate for channel variabilities, as only speaker variability is considered.

The motivation for this study was to implement a new speaker identification system with a smaller I-vector dimension, which cannot be built with normal dimensions using limited speakers for one database. However, different I-vector dimensions can be created using fusion methods to improve system performance. In addition, the main aim of the current system is to provide benchmark evaluations for other researchers working on speaker identification. Furthermore, this study demonstrates how far SIA is affected by the following: I-vector dimension; environment, including the handset and AWGN and various types of NSN; database type; fusion type; and, different I-vector combinations based on four features without fusion.

This chapter is structured as follows: Section 6.1 gives an overview of the background; Section 6.2 focuses on previous work related to the current research; Section 6.3 explains the fundamental framework for speaker identification, including all methodologies for I-vector schemes with fusion to the text independent speaker identification; Section 6.4 describes the simulation setup, including databases and environments; Section 6.5 presents the experimental results, and then elaborates on and analyses these with the discussion; Section 6.6 shows the recent works related to I-vector and GMM-UBM techniques for speaker identification; Section 6.7 presents a summary, including the conclusion; then Appendix 6.1 presents twenty one tables of results for the four databases.

## 6.2 Related Work

This section can be divided into two main parts; the first part is the related work on speaker identification, while the second includes the related work on I-vector speaker identification. In the context of the first part, some researchers extended their aims to include text-independent work on both identification and verification recognition. They applied a non-linear frame likelihood transformation for likelihood normalization purposes as well as likelihood normalization at the frame level, as well as weighted model rank transformation [142]. This system was evaluated for TIMIT and NTT databases, but this work was limited by the number of Gaussian mixtures and did not test different environment conditions. Furthermore, other researchers have applied second order statistics measurements for closed set text independent speaker identification for three different databases, originally derived from the TIMIT database: TIMIT (high quality speech), NTIMIT (telephone speech), and FTIMIT (0-4 kHz of TIMIT) [143]. Although, these comparisons of the three databases are significant, this work is still limited by the lack of testing under real noise conditions.

Additionally [39] developed four databases to evaluate the Gaussian mixture models for both verification and identification recognition in these databases: TIMIT, NTIMIT, Switchboard and YOHO. The main drawback, however, was the limited number of Gaussian mixtures, which may have affected the system accuracy. Moreover, concerning text-independent speaker identification, [144] investigated the relationship between speaker characteristics and frequency components, and tested these with the NTT-VR database. However, this system was not examined for noise robustness. Recent researchers, such as [1], have suggested fusion scores between the MFCC and the inverse MFCC features to improve performance accuracy. 120 speakers from dialect region one and four from the TIMIT database were tested using GMM, but evaluation in noisy environments was missing. A different view for both identification and verification was achieved by [145], which exploited lip biometrics as new features to capture both behavioural and physiological aspects. This was based on joining hierarchical pooling and spatiotemporal sparse coding, for one to 40 speakers, together with matching for identification and one to one matching for authentication.

A robust speaker identification approach was suggested in [59], using neural

responses based metrics. System identification with this system was compared with identification results for MFCC, GFCC, and Frequency Domain Linear Prediction (FDLP) features. Four databases were evaluated, and three of these were text-independent (TIMIT, TIDIGT and YOHO). UM was used for text-dependent with white Gaussian noise, street, and pink background noise. Also, the system was compared with a GMM-UBM approach with 128 Gaussian mixture components. However, this study does not give a balanced comparison of the number of speakers and the sampling rate with limited mixture components in the GMM.

In [24], novel features were investigated by combining the spectrogram for both the Radon Transform (RT) and DCT. The system was tested with the TIMIT and SGGS databases. Although white noise was tested, both handset effects and non-stationary noise were missing from the evaluation. Furthermore, [41] presented a GMM method for text independent speaker identification through noisy telephone channels, and likewise other researchers have emphasised robust speaker identification under noisy environments, for instance: [52], [53], [146] and [147]. Additionally, original speech recordings were evaluated for language and speaker identification tasks in [28] and [61], while in [148] an I-vector was applied for robust language identification and verification recognition. Moreover, different feature modelling approaches were developed for original speech recordings to identify speakers in [149], [29] and [30]. Other researchers employed different approaches for speaker identification, such as employing neural network, wavelet transform, and audio visual approaches and other real time applications [150–154]. In the second part of this section, the related work is considered on speaker identification using an I-vector scheme. In [46] an I-vector approach was investigated with session compensations using LDA, NAP and WCCN for TISI purposes. In spite of 50 self-collected speakers being tested with different classifiers such as SVM and CDS, the study still lacks a sufficient number of tested speakers. For large populations, 1,000 speakers were selected from YouTube for speaker identification using an I-vector framework. On the other hand, a standard database was not used and the evaluation was not under noise conditions [47]. Fusion I-vector and score fusion were also developed by [45] for Speaker Identification (SID), whereas GMM and Hidden Markov Model (HMM) were both used to extract the I-vector. The system complexity was expected to be higher

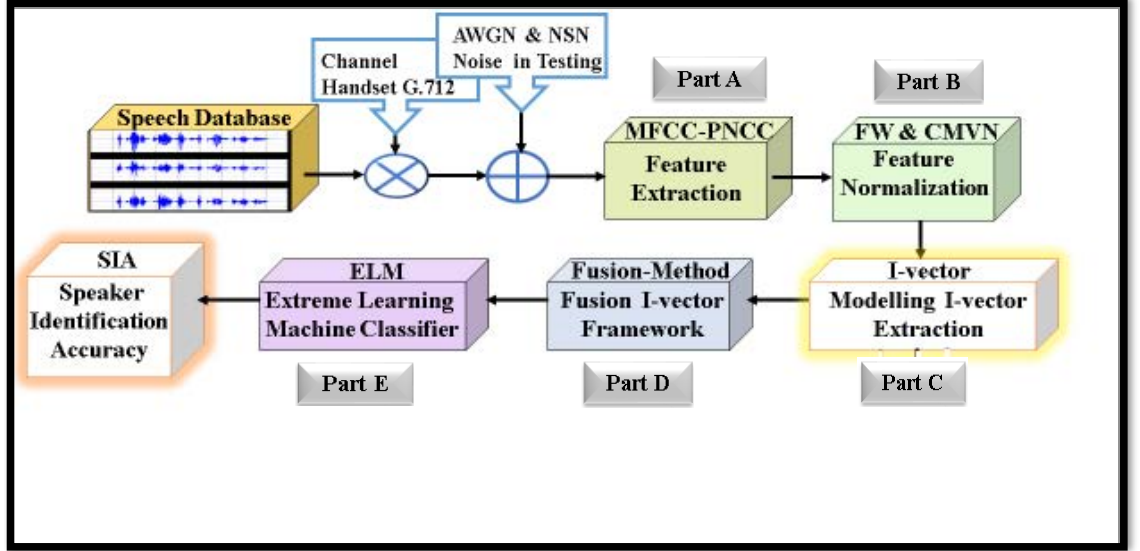


Figure 6.1: Text-Independent Speaker Identification Scheme

than other systems for SID. Moreover, open-set text-independent speaker identification and recognition were evaluated with noisy NIST 2008. This system compared the normalized GMM-UBM with I-vector for different background noise types [48]. Recently, the NIST 2010 was evaluated with a deep neural network, which provides Bottleneck Neural (BN) features and concatenated MFCC features to the I-vector framework, and noisy environments were taken into consideration [49].

### 6.3 Fusion-based I-vector scheme

Fig. 6.1 illustrates the main block diagram for text independent closed-set speaker identification. The system can be divided into five main parts, namely: A, B, C, D and E. In addition, G.712 type handset type at 16 kHz was applied to a normalized speech signal in both training and testing phases, to handle the handset problem effects in this work. Various background noises were added in the testing phase. Basically, the system consists of the following parts: Part A deals with the feature extraction methods (MFCC and PNCC); Part B includes the normalization methods represented by CMVN and FW; Part C involves the modelling system using the I-vector based on the Total Variability Space (TVS), the UBM and the BWS; Part D includes the four I-vector combinations without fusion with d-dimensions,



depending on MFCC and PNCC features (FWMFCC, CMVNMFCC, FWPNCC and CMVNPNCC), and these were fused utilizing seven methods to improve the SIA. The seven fusion methods are: Maximum, Mean, Cumulative, and Weighted sum with  $d$ -dimensional I-vector. In addition, Concatenated and Interleaved fusion with double I-vector dimensions ( $2-d$ ) were also considered, alongside quadruple I-vector dimensions ( $4-d$ ) using concatenation of all MFCC and PNCC feature combinations. Finally, in Part E, the ELM is exploited for signal classification to identify the genuine speakers, and then the speaker identification accuracy is calculated.

#### 6.3.1 Compact features extraction and normalization

In any recognition tasks such as image and pattern recognition, speech and speaker recognition, the extraction of features is the first step. For speaker identification, compact features in the speech data are extracted to attain a good representation of the acoustic signal. In this study, 16 features per speech frame were extracted from both PNCC and MFCC features, and both are represented in Part A of Fig. 6.1. In addition, the features were normalized by employing the FW to mitigate the linear channel mismatch effects, and CMVN to remove the linear channel effects [20]. Both features are depicted in Part B for Fig. 6.1. Moreover, four combinations are investigated and the main infrastructure for these combinations is the MFCC and the PNCC features to produce four feature combinations: FWMFCC, CMVNMFCC, FWPNCC, CMVNPNCC. These are described in [76, 101].

#### 6.3.2 I-vectors extraction framework

The total factors, intermediate vector or i-vector is a compact, fixed length, low dimension and modern state of the art approach. The I-vector is exploited to represent the acoustic models for the MFCC and PNCC feature vectors in order to possibly improve the speaker identification system. The extraction of I-vector is illustrated in Fig. 6.1 as Part C. Substantially, the I-vector can be extracted based on the UBM, which employed the Expectation Maximization (EM) algorithm for training speech utterances to the UBM, TVS and sufficient statistics from centered first order statistics, derived from the BWS algorithm, as shown in Fig. 6.2. Six I-vectors were found from the training data for each speaker, and four I-vectors were developed for testing purposes, according to the training and testing strategy

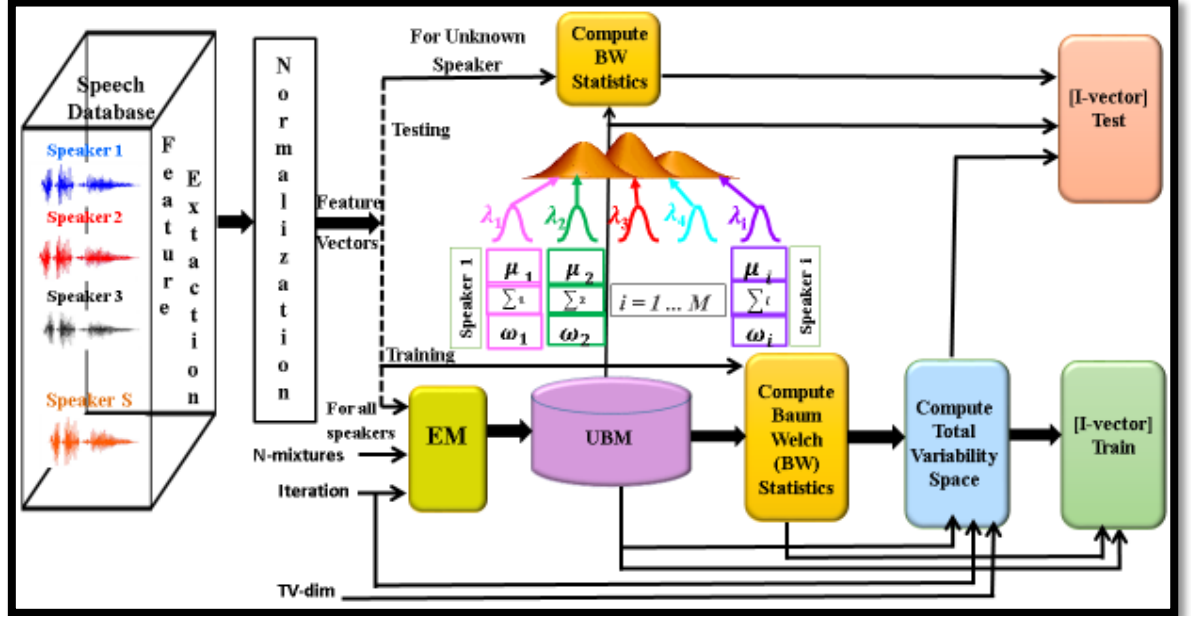


Figure 6.2: I-vector Extraction Block Diagram

for the previous work in [76], in order to accomplish a fair comparison. In [119], the I-vector was initially employed, which gave fixed and low dimension for speaker verification whilst, this work exploited the I-vector for identification purposes. The mathematical framework to construct the I-vector is listed in [33], [119] and [36] and the important equations used are (6.1), (6.2) and (6.3).

$$\mathbf{S} = \boldsymbol{\mu} + \mathbf{U}\mathbf{x} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z} \quad (6.1)$$

$$\mathbf{S} = \boldsymbol{\mu} + \mathbf{T}_V \mathbf{i} \quad (6.2)$$

where:  $\mathbf{S}$  is a dependent supervector for both speaker and channel;  $\boldsymbol{\mu}$  is the independent supervector for the speaker and channel;  $\mathbf{V}$ ,  $\mathbf{U}$  and  $\mathbf{D}$  are the speaker, channel and diagonal residual matrices, respectively. In addition, factors  $\mathbf{y}$ ,  $\mathbf{x}$  and  $\mathbf{z}$  represent the speaker, channel and residual factor, respectively. Furthermore, the low rank matrix is called the total variability  $\mathbf{T}_V$ , while the total factor identity vector is  $\mathbf{i}$ . Finally, I-vectors can be determined as in [119], and described in ((6.3):

$$\mathbf{i} = (\mathbf{I} + (\mathbf{T}_V)^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{N}}(\mathbf{u}) \mathbf{T}_V)^{-1} (\mathbf{T}_V)^T \boldsymbol{\Sigma}^{-1} \check{\mathbf{F}}(\mathbf{u}) \quad (6.3)$$

### 6.3 Fusion-based I-vector scheme

---

where:  $\mathbf{i}$  is the identity vector (I-vector),  $\mathbf{T}_V$  is the total variability matrix,  $\mathbf{I}$  is the identity matrix,  $\mathbf{\Sigma}$  is the  $(CF \times CF)$  diagonal covariance matrix, where  $C$  is the number of mixture components,  $F$  is the dimension of the feature vectors. Furthermore,  $\mathbf{u}$  is the given speech utterance and  $\hat{\mathbf{N}}(\mathbf{u})$  is a diagonal matrix of dimension  $(CF \times CF)$ ,  $\tilde{\mathbf{F}}$  is the  $(CF \times 1)$  dimension supervector and is obtained by concatenating all first-order BaumWelch statistics, and  $(.)^T$  denotes transpose. To summarize the main steps to extract the I-vector are as follows [20]:

Step 1: Forming a UBM from training data using the EM algorithm and Gaussian mixture components for the speakers.

Step 2: Extract the sufficient statistics for the training features using the Baum-Welch (BW) algorithm.

Step 3: Learning a total variability subspace.

Step 4: Extract the I-vector.

#### 6.3.3 Fusion Methods Based on I-vectors

Fig. 6.1 Part D represents the fusion of normalized MFCC I-vectors (FWMFCC and CMVNMFCC) with the corresponding normalized PNCC I-vectors (FWPNCC and CMVNPNC).

According to Fig. 6.3, seven I-vector fusions are achieved, namely: Maximum, Mean, Cumulative and Weighted sum (with  $d$ -dimension), Concatenated and Interleaving fusion (with  $2d$ -dimension), and concatenated with  $4d$ -dimension. Four out of seven fusion techniques are employed with the same input I-vector dimension, but only interleaving and concatenating fusion give double the input I-vector dimension. Essentially, normalized PNCC I-vector features ( FWPNCC and CMVNPNC ) with  $d$ -dimension are fused with the corresponding  $d$ -dimension of the normalized MFCC I-vector features ( FWMFCC and CMVNMFCC ). However, all fusion approaches give the same  $d$ -dimension as a consequence of the fusion process. In contrast, both interleaved and concatenated fusion produce double dimension ( $2d$ ), and, finally the concatenated fusion of all feature combinations of I-vectors produces  $4d$  dimension. In addition, through the I-vectors fusion process, the training I-vectors based on MFCC features are fused with the corresponding training I-vectors from PNCC features. The process is similarly achieved for testing I-vectors. The seven fusion-based I-vectors

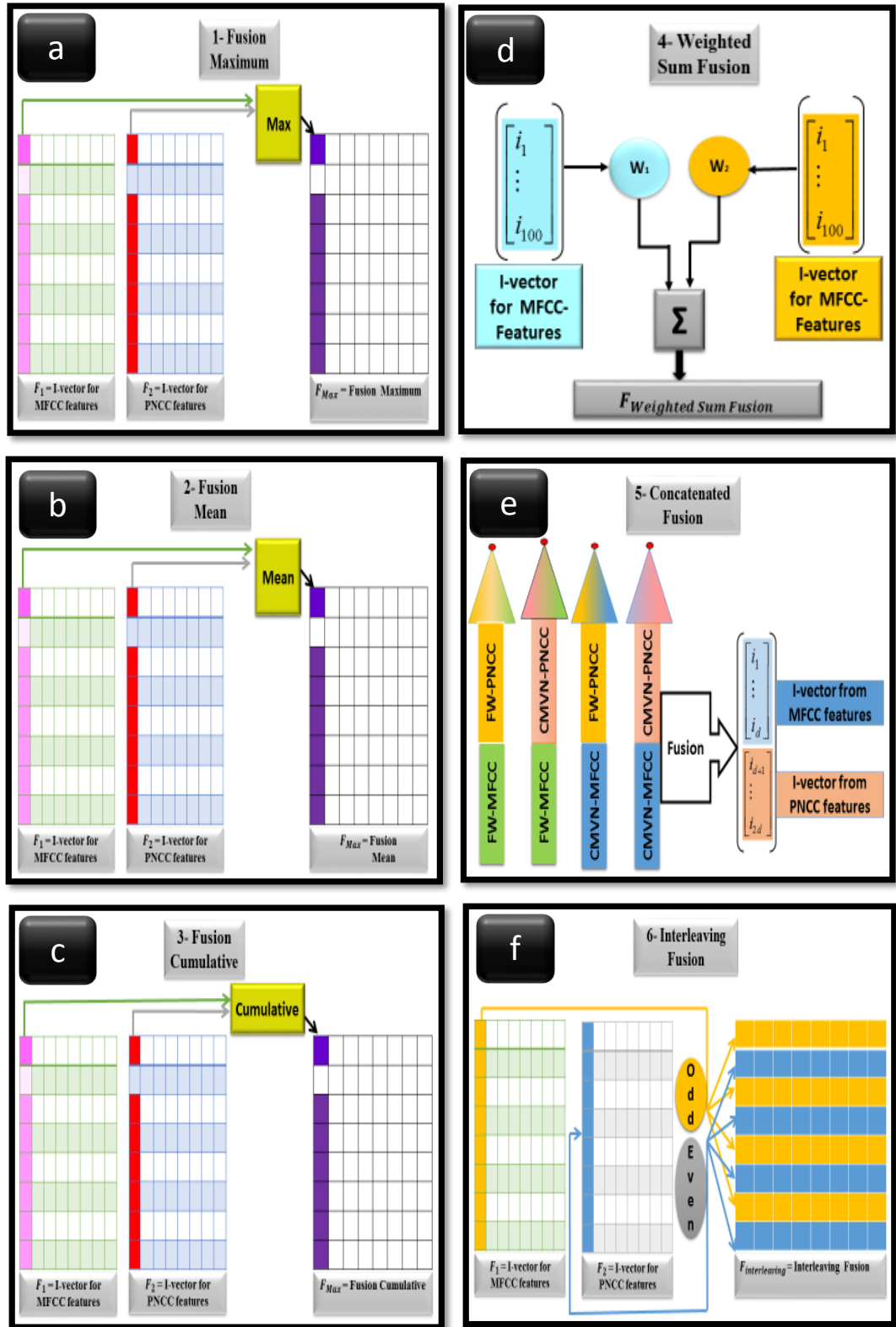


Figure 6.3: I-vector Fusion Scheme Methods: (a) Maximum, (b) Mean, (c) Cumulative and (d) Weighted Sum (with  $d$ -Dimension), (e) Concatenated with  $2d$  and  $4d$ -Dimensions, and (f) Interleaving Fusion (with  $2d$ -Dimension)

### 6.3 Fusion-based I-vector scheme

---

techniques are explained below in equations (6.4)-(6.10):

$$\mathbf{i}_{WSF} = \omega_k \dot{\mathbf{i}}_j + (1 - \omega_k) \ddot{\mathbf{i}}_j \quad (6.4)$$

where:  $k = 1, 2, 3, 4$ . while,  $\omega_1, \omega_2, \omega_3$  and  $\omega_4 = 0.9, 0.8, 0.77$  and  $0.7$  respectively, which have been found to yield a higher identification rate empirically.

$$\mathbf{i}_{Maximum} = \max(\dot{\mathbf{i}}_j, \ddot{\mathbf{i}}_j) \quad (6.5)$$

$$\mathbf{i}_{Mean} = (\dot{\mathbf{i}}_j + \ddot{\mathbf{i}}_j)/2 \quad (6.6)$$

$$\mathbf{i}_{Cumulative} = \dot{\mathbf{i}}_j + \ddot{\mathbf{i}}_j \quad (6.7)$$

$$\mathbf{i}_{Concatenated(2d)} = \begin{bmatrix} \dot{\mathbf{i}}_j^T & \ddot{\mathbf{i}}_j^T \end{bmatrix}^T \quad (6.8)$$

$$\mathbf{i}_{interleaving(2d)} = \begin{bmatrix} \dot{\mathbf{i}}_{jj_{odd}}^T & \ddot{\mathbf{i}}_{jj_{even}}^T \end{bmatrix}^T \quad (6.9)$$

$$\mathbf{i}_{Concatenated(4d)} = \begin{bmatrix} \dot{\mathbf{i}}_{M1}^T & \dot{\mathbf{i}}_{M2}^T & \dot{\mathbf{i}}_{P1}^T & \dot{\mathbf{i}}_{P2}^T \end{bmatrix}^T \quad (6.10)$$

and  $d$  = dimension of I-vector, where  $j = 1, \dots, d$ ,  $r = d+1, \dots, 2d$ ,  $jj = 1, \dots, 2d$ ,  $\dot{\mathbf{i}}$  is the I-vector for the normalized MFCC features, which has the highest SIA of CMVNMFCF and FWMFCF, which are denoted by  $\mathbf{i}_{M1}$  and  $\mathbf{i}_{M2}$ ,  $\ddot{\mathbf{i}}$  is the normalized PNCC I-vector features, which has the highest SIA of FWPNC and CMVNPNC, denoted by  $\mathbf{i}_{P1}$  and  $\mathbf{i}_{P2}$ .  $\mathbf{i}_{WSF}$ ,  $\mathbf{i}_{Maximum}$  and  $\mathbf{i}_{Mean}$  are the weighted sum, maximum and mean fusion I-vectors with  $d$ -dimensions. Also,  $\mathbf{i}_{Cumulative}$  is the Cumulative fusion with the  $d$ -dimension I-vector;  $\mathbf{i}_{Concatenated(2d)}$  and  $\mathbf{i}_{interleaving}$  are Concatenated and Interleaved fusion I-vectors with  $2d$ -dimension I-vector;  $\mathbf{i}_{Concatenated(4d)}$  is the concatenated fusion I-vectors for all feature combinations of I-vector (without fusion) with  $4d$ -dimension for all feature combinations of the I-vectors.

### 6.3.4 ELM classification and calculating the identification accuracy

Recently, ELMs have received significant interest from various research fields. They have been used in widespread areas such as computer vision, biomedical engineering and control and robotics, because of their simple, efficient and impressive performance [155]. ELM operates as a single hidden layer feed-forward network, in which the hidden layer does not need to be tuned and no iterations are required. The connections of these hidden neurons or nodes are randomly generated independently from the training dataset or the target functions. ELM is used for extremely fast learning, and classification and regression applications. In addition, the ELM has the following properties: this fast method saves time; testing accuracy is high; it is independent of both training dataset and target function; before the training data are seen ELM randomly generates the connections of the hidden nodes.

In this study, an ELM was employed, as explained in Fig. 6.1 Part E, as a single layer feedforward network classifier to identify genuine speakers. The ELM was exploited for speaker identification [61]. Fig. 6.4 shows the ELM network scheme also produced in [156] and [157]. The number of input nodes is equal to the I-vector dimension, which is equal to the number of hidden neurons in most of the experiments used in this work. In addition, different neuron numbers are also selected when it is necessary to achieve higher performance accuracy. However, the number of output neurons is equal to the number of classes, and this work developed 120 classes to represent 120 speakers. The activation function is also required, and sigmoid function is used in this chapter. Furthermore, details about the ELM can be found in [158], [159], [160], [161] and [162]. To calculate the output weights of the ELM, the following equations are used to find regularized values [156] [159].

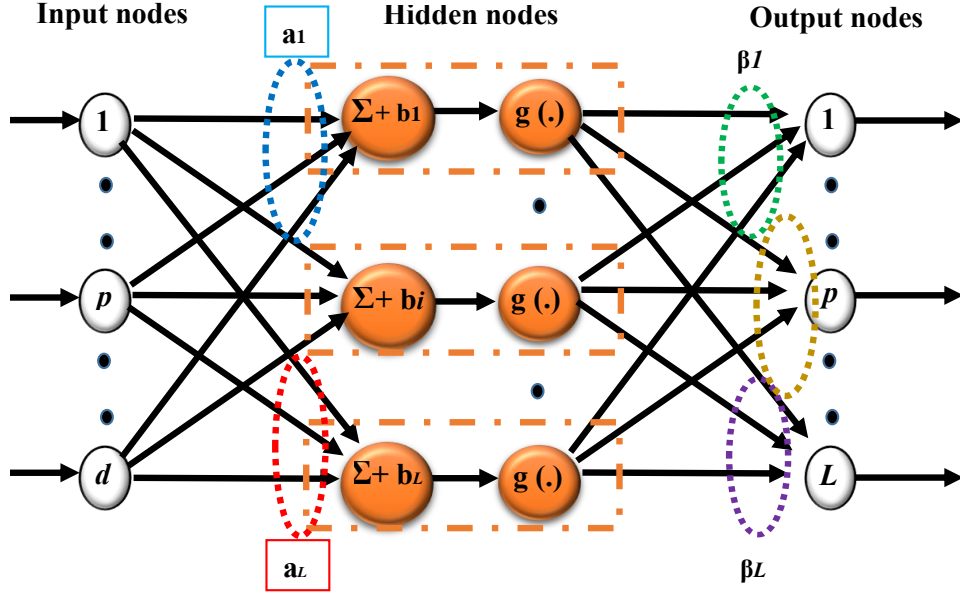


Figure 6.4: Structure of Single Layer Feedforward Extreme Learning Machine with Input Dimension  $d$ ,  $L$  hidden nodes and  $L$  outputs [159]

$$\mathbf{H}\boldsymbol{\beta} = \check{\mathbf{T}} \quad (6.11)$$

$$\boldsymbol{\beta} = \mathbf{H}^\dagger \check{\mathbf{T}} \quad (6.12)$$

$$\boldsymbol{\beta} = \left( \frac{\mathbf{I}}{r} + \mathbf{H}^T \mathbf{H} \right)^{-1} \mathbf{H}^T \check{\mathbf{T}} \quad (6.13)$$

where:  $\boldsymbol{\beta}$  is the output weight matrix  $\boldsymbol{\beta} = [\beta_1, \dots, \beta_L]^T$ ,  $\mathbf{H} = [h^T(\mathbf{x}_1), \dots, h^T(\mathbf{x}_N)]^T$ ,  $h(\mathbf{x}) = [g(a_1^T \mathbf{x} + b_1), \dots, g(a_L^T \mathbf{x} + b_L)]$  is the hidden node outputs,  $\mathbf{x}$  is the input vector,  $g(a_j^T \mathbf{x} + b_j)$  is the output of the  $j^{th}$  hidden node.  $N$  is the number of training samples,  $\check{\mathbf{T}}$  is the target label matrix,  $\check{\mathbf{T}} = [\check{t}_1, \dots, \check{t}_N]^T$ ,  $\mathbf{H}^\dagger$  is the Moore-Penrose generalized inverse of matrix  $\mathbf{H}$ ,  $r$  is the regularization factor. The target label matrix  $\check{\mathbf{T}}$  is represented by (no. of speakers  $\times$  no. of training samples ( $L \times N$ )) dimensions. Each I-vector with 100 dimension is entered into the ELM classifier. The outputs represent the speaker classes in this work of 120 speakers. In addition, the actual outputs are real numbers, while the maximum is selected for the output vector and this maximum refers to the identified speaker position. Ultimately, the SIA is as determined in equation 3.4.

## 6.4 Simulation Setups

---

To summarize the main steps for ELM classifier then calculation the SIA as follows:

Step 1: Compute the I-vectors from training and testing data.

Step 2: Assume the number of output neurons is equal to the number of speakers.

Step 3: Form the target matrix for training.

Step 4: Generate randomly the input weights and biases for the hidden neurons.

Step 5: Suppose the number of hidden neurons is equal to the I-vector dimension.

Step 6: Calculate hidden neurons output Matrix H.

Step 7: Calculate the output weights.

Step 8: Apply the testing data to the trained ELM.

Step 9: Calculate the genuine speaker identified for each training example by the position of the maximum in the output of the ELM.

Step 10: Calculate the Speaker Identification Accuracy.

In addition, the fusion based I-vectors can be achieved in step 1 by fusing the the I-vectors found with MFCC and PNCC features.

## 6.4 Simulation Setups

Table 6.1 shows the parameters used in all simulations used for this chapter, which involves four databases, as well as system details, background noise and various challenging environments. This table also includes the number of speakers, training and testing partitioning for each database, feature types, normalization methods, classifier method and the fusion techniques.

### 6.4.1 Databases and Environments

This chapter considers four main databases: TIMIT Acoustic-Phonetic Continuous Speech Corpus-1993 [65]; the 2016 SITW Speaker Recognition Challenge [79] and [80]; the 2008 NIST Speaker Recognition Evaluation Training Set Part 2-2011 [84]; NTIMIT, available on [66]. However, the AWGN and NSN were only in the testing phase with seven SNR levels (0dB to 30dB) with step size 5dB for each level, based on the corresponding noise power, as in [76]. The NSN is available from [86] and [85], which were both used in the testing phase. In addition, A G.712 type handset at 16 kHz with a 4<sup>th</sup> order linear IIR filter was derived from the Z transform multiplication



## 6.4 Simulation Setups

Table 6.1: Parameters and Setup Used in All Experiments and Simulations

Aspects	Parameters and Experimental Setup
Sampling frequency	16000
Window type	Hamming
Frame length	16 ms
Frame shift	8 ms
Pre-emphasis factor	0.96
Databases	<b>TIMIT</b> , <b>SITW</b> and <b>NIST 2008</b> , <b>NTIMIT</b>
Number of speakers	120 speakers for each database, total 480 speakers for all databases
Total speech utterances used	1,200 for each database, total 4,800 for all databases
Language	English
Data Source (s)	Microphone Speech for TIMIT and NIST 2008, Hand Annotated Speech from Open Source Media for SITW, Telephone Speech for NTIMIT
No. of samples per speaker	10 for each of TIMIT and NTIMIT, 10 created as well for both SITW and NIST 2008
Testing samples for each database	Total 480 utterances
Training samples for each database	Total 720 utterances
Dialect region	In this chapter, 49 speakers from DR1 & 71 from DR4 are selected for each of TIMIT and NTIMIT databases
Average sample duration	8 seconds in length 129250 (for each speech utterance in both training and testing); All speech samples were taken with fixed length of 129250 samples; concatenation is applied where necessary
Features	MFCC and PNCC
Features dimension	16
Feature normalization	Feature warping (FW) and Cepstral Mean Variance Normalization (CMVN)
Modelling	I-vector
Classifier	Extreme Learning Machine (ELM)
UBM Mixture Sizes	{8, 16, 32, 64, 128, 256, 512}
Fusion Types	Fusion I-vectors methods: Mean, Weighted Sum, Maximum, Cumulative $d$ -dimension Concatenated, Interleaved $2d$ -dimension Concatenated $4d$ -dimension
System Environment	Original speech recordings, AWGN with G.712 type handset at 16 kHz and (Street-traffic, Bus-interior and Crowd talking NSN) with handset
SNR levels in dB	{0, 5, 10, 15, 20, 25, 30}

of two second order cascaded filters, as previously exploited in [3]. In this chapter, the G.712 type handset at 16 kHz is applied to a normalized speech signal for both training and testing phases, as employed in [76]. Furthermore, all noise, handset and databases, except for the NTIMIT database, are explained in more depth in Chapters 3 and 5.

## 6.5 Experimental Results and Discussions

In this chapter, according to the databases used, four main experiments are considered, namely Experiments 1-4. The results for all experiments are shown in Appendix 6.1, from Table 6.3 to Table 6.23. Table 6.3 - Table 6.12 show the SIA results for the TIMIT database, while Table 6.13 - Table 6.17 give those for the SITW database. Table 6.18 - Table 6.22 illustrate the results for the NIST 2008 database, and Table 6.23 shows the single simulation for the NTIMIT database.

In Experiment 1, there were ten simulations and their results are represented in Table 6.3-Table 6.12, for the 120 speakers (1,200 speech utterances) from the TIMIT database, based on four combinations of I-vectors without fusion, and seven different I-vector fusion methods. Table 6.3 shows the SIA for original speech recordings with the TIMIT database, based on 100, 200 and 400 I-vector dimensions, while Table 6.4 considered 200, 400 and 800 dimensions. Table 6.5 - Table 6.8 show the SIA under AWGN, street traffic NSN, bus interior NSN and crowd talking NSN, respectively, without handset for 100, 200 and 400 I-vector dimensions. However, Table 6.9 - Table 6.12 are the corresponding AWGN, street NSN, bus-interior NSN, crowd talking NSN with G.712 type handset at 16 kHz for the same 100, 200 and 400 I-vector dimensions. In all tables Table 6.3 - Table 6.12, the SIA for the four feature combinations of I-vectors (without fusion) are: FWMFCC, CMVNMFCC, FWPNCC and CMVNPNC, and the SIA of the seven fusion methods are presented in this chapter.

In Experiment 2, there were five simulations for the 120 speakers (1,200 speech utterances) of the SITW database, and the results are presented in Table 6.13 - Table 6.17. These tables show the SIA under the original speech recordings, AWGN, street NSN, bus-interior NSN, and crowd talking NSN with G.712 type handset at 16 kHz based on 100, 200 and 400 I-vector dimensions. Similarly, in Experiment 3, for the same environments, five simulations were completed and their results are given in (Table 6.18-Table 6.22) for the 120 speakers (1,200 speech utterances) from the NIST 2008 database. However, in Experiment 4, only one simulation was completed and the results for this are explained in Table 6.23 for 1,200 speech utterances (120 speakers) with the telephone channel NTIMIT database. According to [163], the SIA was calculated for only 100 I-vector dimension for the four combinations of I-vector (without fusion) and with three fusion methods (weighted sum, maximum

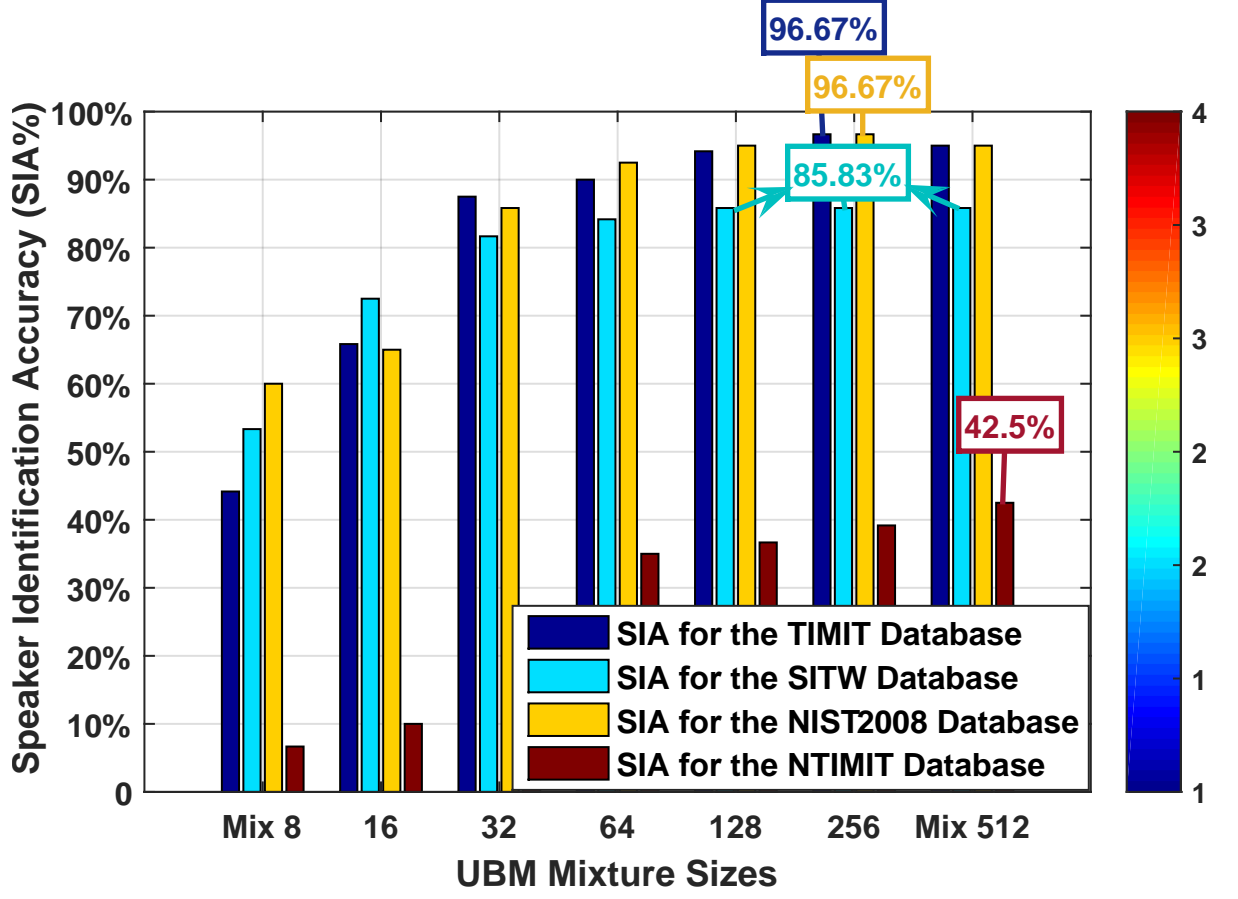


Figure 6.5: The Highest SIA Using the I-vector with 100, 200 and 400 Dimensions per UBM Mixture Size for TIMIT, SITW, NIST2008 and NTIMIT Databases; which Represented the Best SIA for the Tables: Table 6.3, Table 6.13, Table 6.18 and Table 6.23, Respectively

and mean fusion). The system was evaluated using 120 speakers from each of TIMIT and NTIMIT databases. The results were taken from Table 6.3 and Table 6.23.

The analysis and discussion for all these results is complex. Therefore, in this chapter the best SIA was selected for each mixture or each SNR level, according to similar environment for all databases, regardless of feature combination type (without fusion) or fusion method and/or the I-vector dimensions. Then, these results are presented in figures so that they can be analysed and discussed clearly and easily.

Firstly, according to Fig. 6.5, the best SIA using the I-vector with 100, 200 and 400 dimensions using the original speech recordings from four databases and are given as follows: Table 6.3 for TIMIT; Table 6.13 for SITW; Table 6.18 for NIST 2008, and Table 6.23 for NTIMIT. These are presented based on the UBM mixture

## 6.5 Experimental Results and Discussions

---

sizes  $\{8, 16, 32, 64, 128, 256, 512\}$ . However, from this figure, it can be observed that increasing UBM mixture sizes also increases the SIA for all databases. The highest SIAs were 96.67%, 85.83%, 96.67% at UBM mixture size 256 for the TIMIT, SITW, NIST 2008, respectively. SIA of 42.5% was achieved at a mixture of 512 for the NTIMIT database. It is clear from the figure that the performance for the NTIMIT database was the worst, compared with other databases because of the noise effects on the telephone channel from the NTIMIT database. Therefore, this database was excluded for all other environments and experiments. In addition, the order for these databases, according to the best to worst SIA, is as follows: NIST 2008, TIMIT, SITW and NTIMIT. It clear that the SIA for NIST 2008 database is better than the SIA for the TIMIT database at UBM mixture sizes  $\{8, 64, 128\}$ , while TIMIT is best at sizes (16, 32), and both are equal at the mixtures  $\{256 \text{ and } 512\}$  with the same highest SIA 96.67% at the same mixture size 256.

Secondly, Fig. 6.6 depicts the best SIA per SNR level using the I-vector with 100, 200 and 400 dimensions at UBM mixture size 256, under AWGN with the G.712 type handset at 16 kHz (Table 6.9 for TIMIT, Table 6.14 for SITW, Table 6.19 for NIST 2008). In addition, increasing the SNR levels also increased the SIA for all the databases used in this simulation. The highest SIAs were 74.17%, 84.17%, 81.7% at SNR with 30 dB for the TIMIT, SITW and NIST 2008, respectively. It is evident that the new database (SITW) had the best performance compared with other databases. However, the second best SIA was for the NIST 2008; in contrast, TIMIT had the worst performance accuracy. This was expected with noisy speech databases such as the SITW and NIST 2008, which are robust under conditions of background noise compared with an ideal acquisition database, such as TIMIT.

Thirdly, Fig. 6.7 shows the best SIA at each SNR level using the I-vector with 100, 200 and 400 dimensions at UBM mixture size 256 for street traffic with the G.712 type handset at 16 kHz (Table 6.10 for TIMIT, Table 6.15 for SITW, Table 6.20 for NIST 2008). The highest SIAs were 82%, 84.17%, 78.33% at SNR 30 dB for TIMIT, SITW and NIST 2008, respectively. Likewise with Fig. 6.6, it is clear that for street NSN the SITW had the best performance, compared with other databases, the second best SIA was the TIMIT database, and the worst was the NIST 2008 database. These results were due to the non-stationary characteristics of the street noise. Furthermore, increasing the SNR increases the SIA for all the databases in this figure.

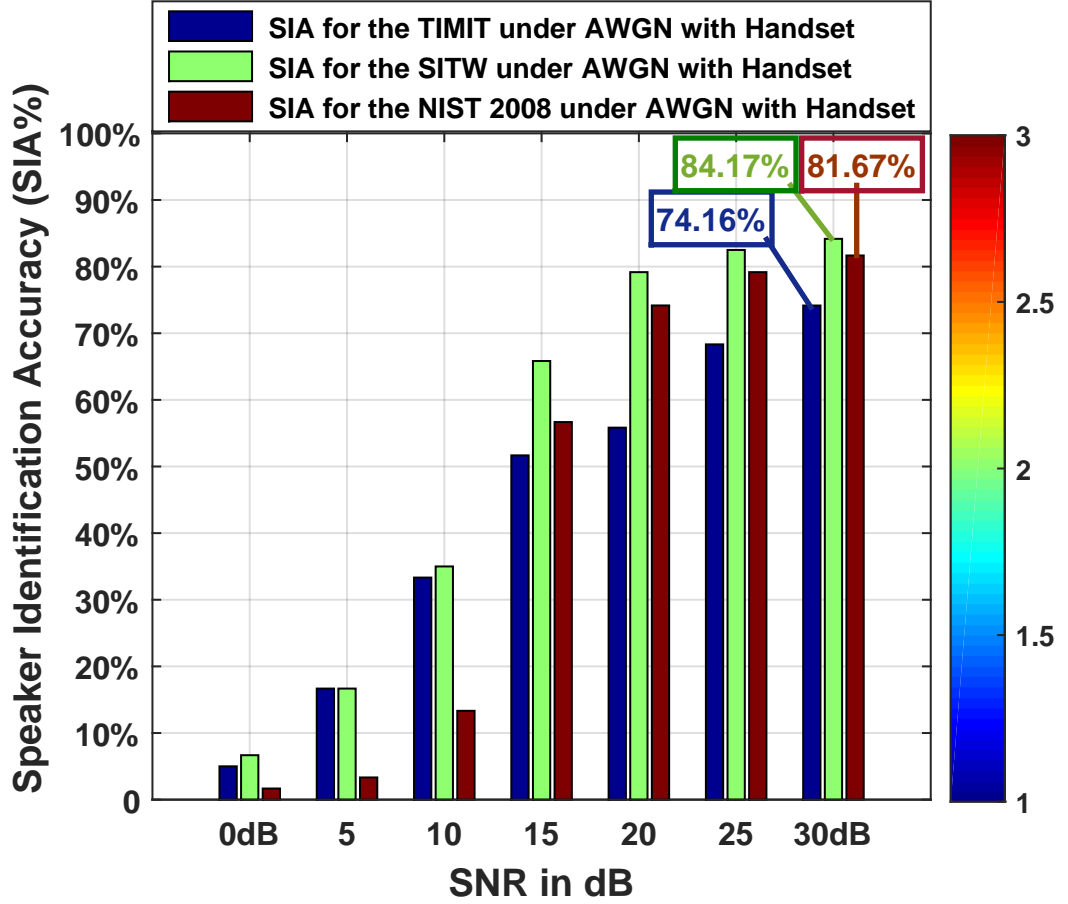


Figure 6.6: The Bar Chart Shows the Highest SIA at each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT, SITW and NIST2008 Databases Under AWGN with G.712 Type Handset at 16 kHz; the Best SIAs are found in: Table 6.9, Table 6.14 and Table 6.19, Respectively

Fourthly, in Fig. 6.8 shows the best SIA at each SNR level using the I-vector with 100, 200 and 400 dimensions at UBM mixture size 256 for bus interior NSN with the G.712 type handset at 16 kHz (Table 6.11 for TIMIT; Table 6.16 for SITW; Table 6.21 for NIST 2008). The highest SIAs were 89.17%, 86.67%, 87.5% at SNR with 30 dB for TIMIT, SITW and NIST 2008, respectively. This figure demonstrates that the best SIA was achieved for TIMIT, then NIST 2008, and SITW. Although the SITW database had the lowest SIA, it was still close to other databases.

Finally, Fig. 6.9 shows the highest SIA for each SNR level using the I-vector with 100, 200 and 400 dimensions at UBM mixture size 256 under crowd talking NSN with the G.712 type handset at 16 kHz (Table 6.12 for TIMIT; Table 6.17 for SITW; Table 6.22 for NIST 2008). The highest SIAs were equal in all databases, with 85% at SNR 30 dB for the TIMIT, SITW, and NIST 2008, respectively. Increasing the

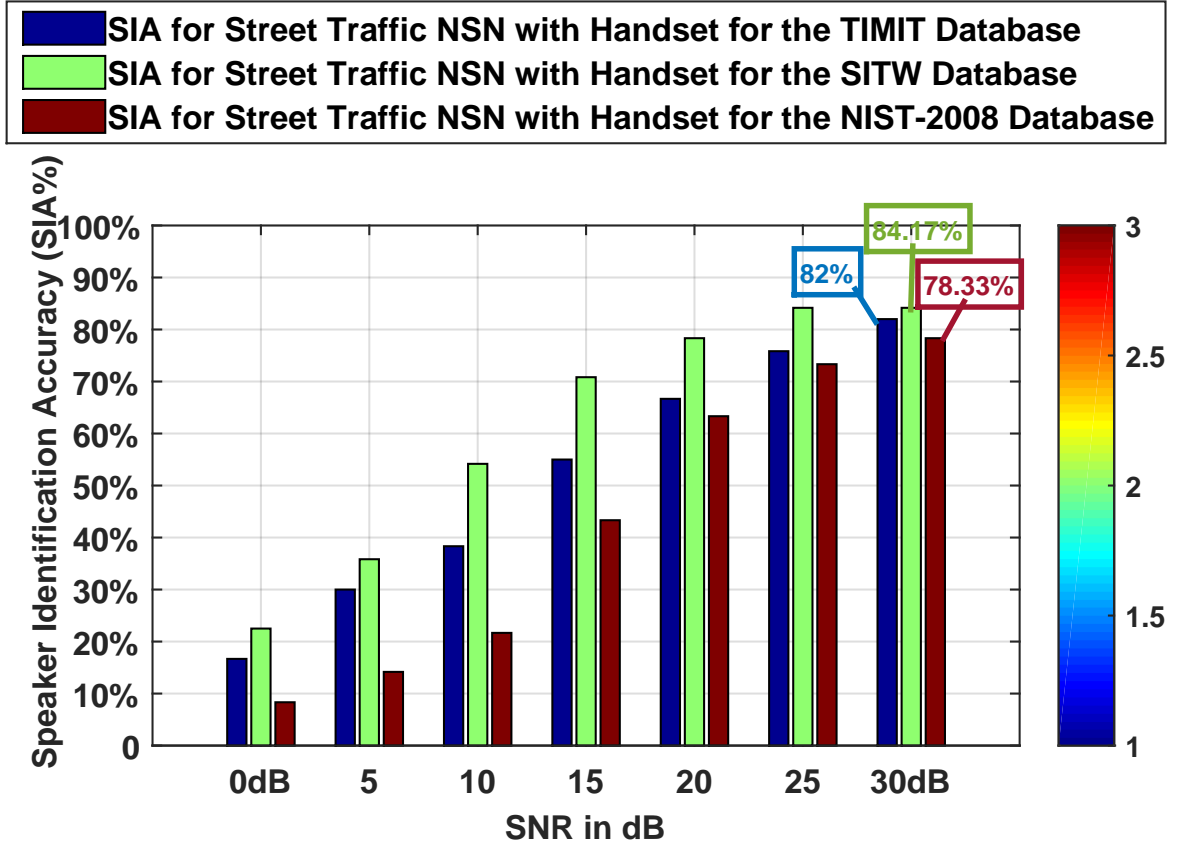


Figure 6.7: The Bar Chart Shows the Highest SIA at Each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT, SITW and NIST2008 Databases Under Street Traffic NSN with G.712 Type Handset at 16 kHz; the Best SIAs are found in: Table 6.10, Table 6.15 and Table 6.20, Respectively

SNR level increased the SIA for all the databases, but it is clear that the SITW performance significantly outperformed both TIMIT and NIST 2008 for SNR levels (0-25) dB, and for (0-15) dB the second best SIA was the TIMIT database. In addition, for SNR levels (20-25) dB, the NIST 2008 databases seem better than the SIA for the corresponding SNR for the TIMIT database. Furthermore, additional evaluations for the TIMIT database only, as seen in Fig. 6.10, show the highest SIA for each SNR level using the I-vector with 100, 200 and 400 dimensions at UBM mixture size 256 with AWGN, street NSN, bus NSN and crowd talking NSN without the handset effect. This figure includes the best SIAs in Table 6.5 to Table 6.8 for the TIMIT database. In this figure, increasing the SNR increases the SIA for all simulations, in all environments, from the TIMIT database. The highest SIAs were 80.83%, 90%, 93.33% and 90.83 % at SNR with 30 dB for the TIMIT database with AWGN, street traffic NSN, bus interior NSN and crowd talking NSN, respectively. It

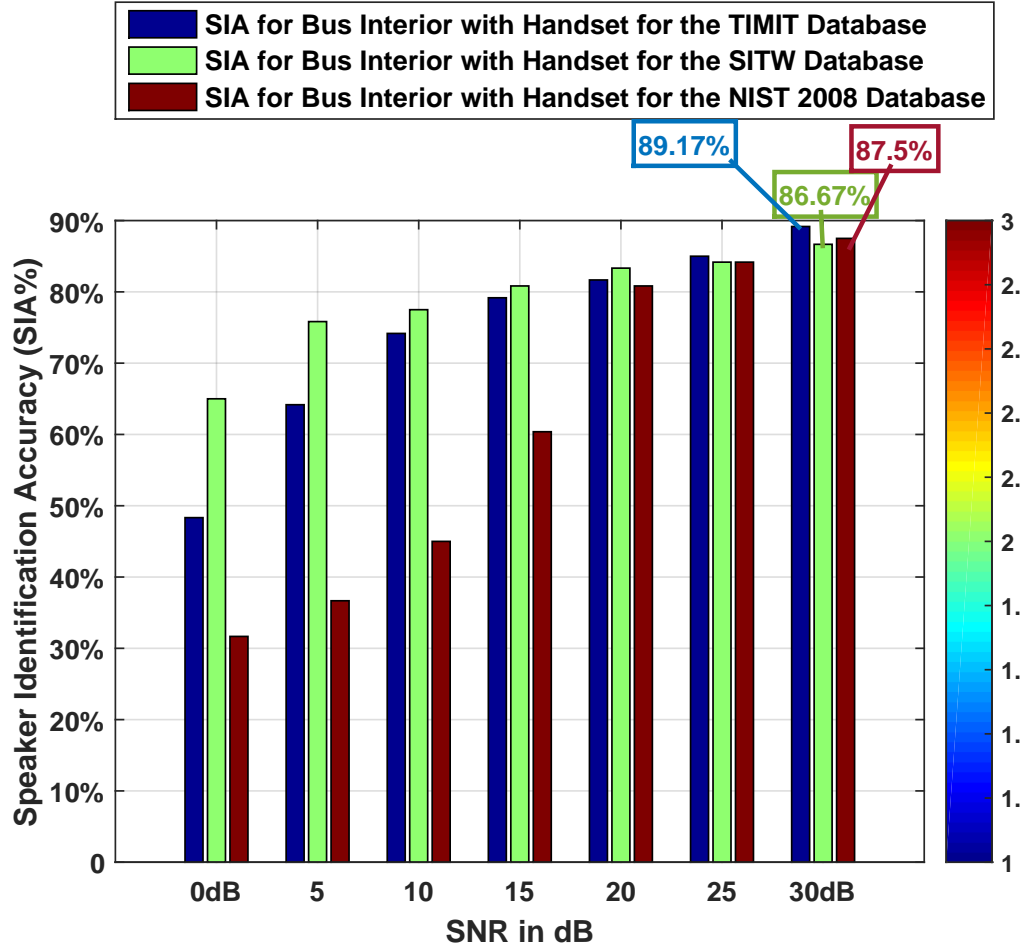


Figure 6.8: The Bar Chart Show the Highest SIA for Each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT, SITW and NIST2008 Databases Under Bus Interior NSN with G.712 Type Handset at 16 kHz; the Best SIAs are found in: Table 6.11, Table 6.16 and Table 6.21, Respectively

is very clear from the figure that the worst results were achieved in the presence of a stationary spectrum of noise (AWGN). In addition, depending on the non stationary noise, which has unequal noise distribution, the results were similar under street and crowd talking NSN. Each outperforms the other with different SNR levels. In contrast, the bar chart under bus interior NSN had the best SIA compared with other types of noise, because of the natural characteristics and spectrum, as discussed in Chapter 5. In addition, this chapter discusses three major questions in three subsections in this part of discussion:  $Q_{6.1}$  the first question concerns how far the I-vector dimensions affect the SIA; the second question is  $Q_{6.2}$  how far the UBM mixture sizes, SNR level, feature combination of I-vector, with and without fusion, affected the SIA; the final question is  $Q_{6.3}$  which is best, GMM-UBM or I-vector

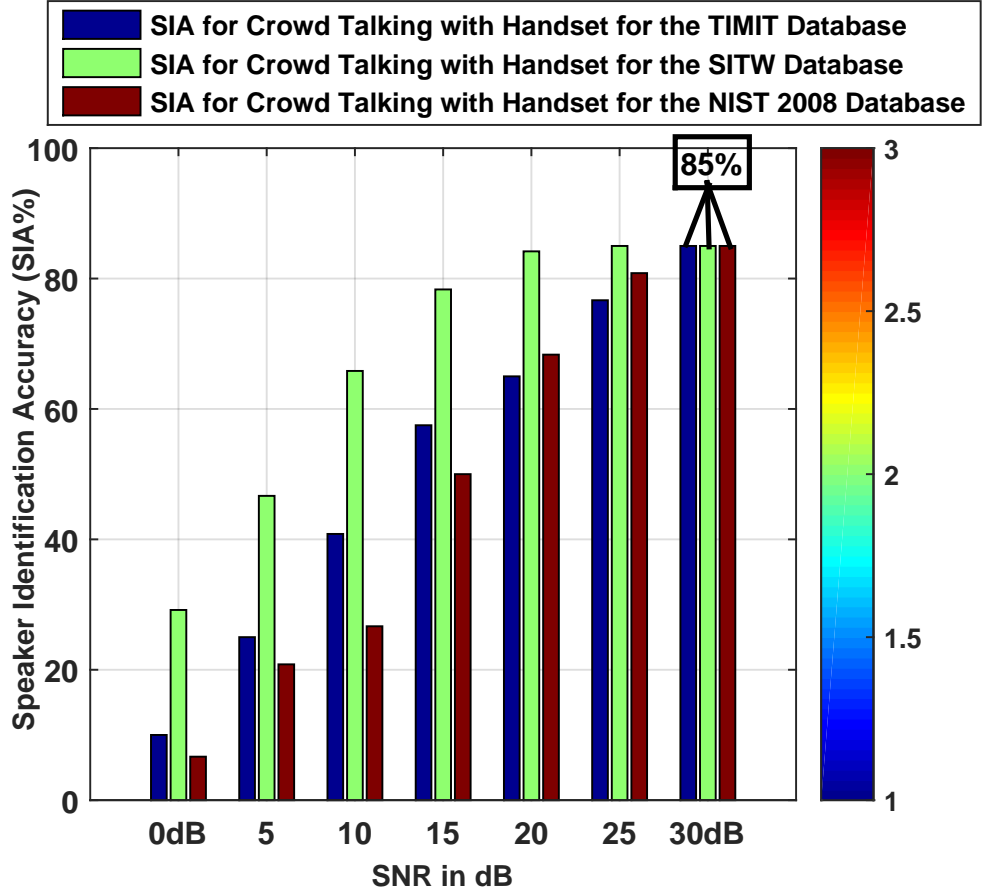


Figure 6.9: The Bar Chart Show the Highest SIA for Each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT, SITW and NIST2008 Databases Under Crowd Talking NSN with G.712 Type Handset at 16 kHz; the Best SIAs are found in: Table 6.12, Table 6.17 and Table 6.22, Respectively

approaches for speaker identification.

### 6.5.1 The Relationship Between Multi-Dimensional I-vectors and SIA in TIMIT Database Evaluations

In order to answer the first question  $Q_{6.1}$ , this subsection considers the best SIA results as presented in Table 6.3 and Table 6.4 which were used to create Fig. 6.11. This gives a good representation of the I-vector dimensions (100, 200, 400 and 800) with wide ranges of UBM mixture sizes against the SIA for original speech recordings for the TIMIT database. It is clear that the best SIA 96.67% was achieved at small I-



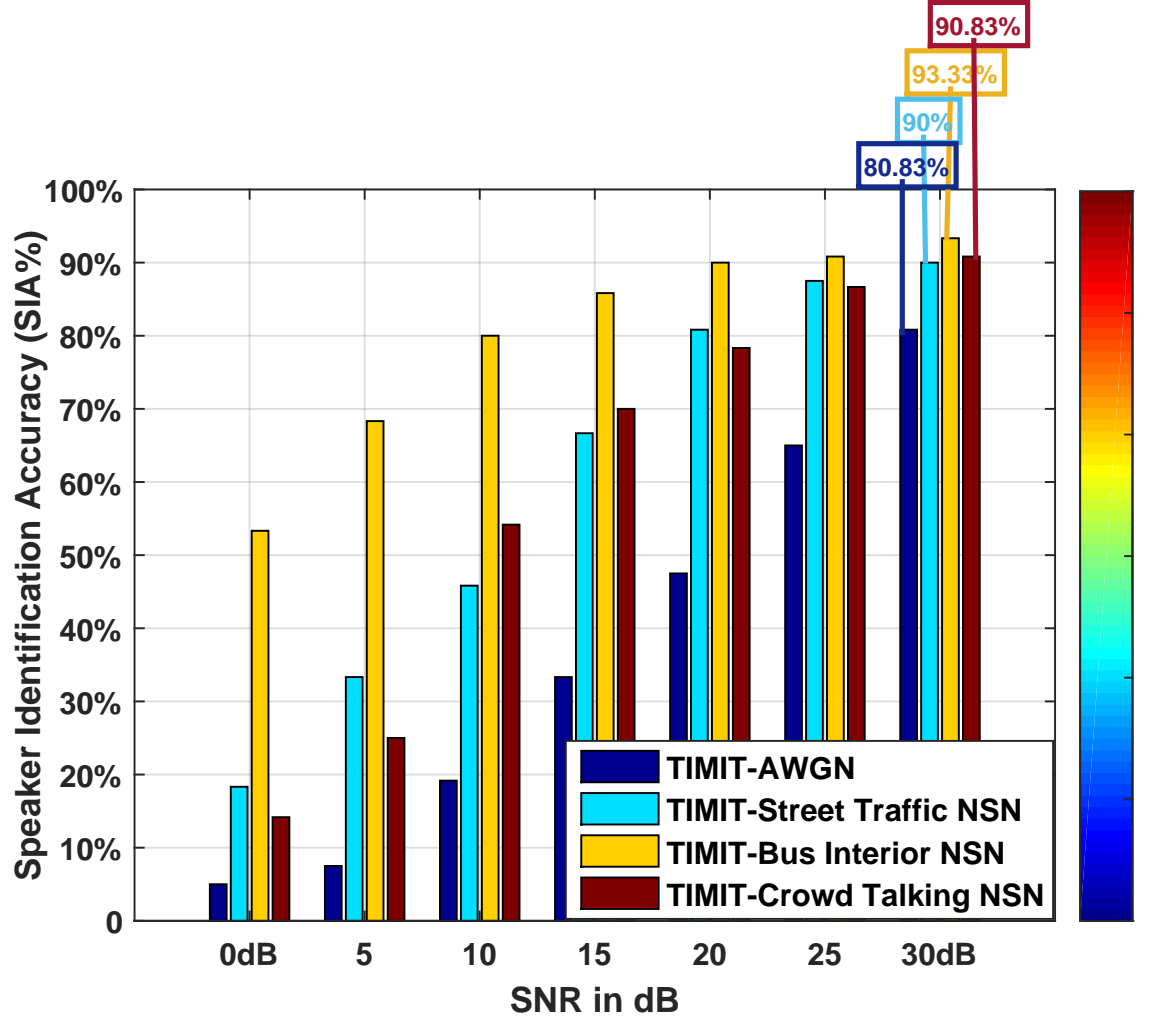


Figure 6.10: The Bar Chart Shows the Highest SIA for Each SNR Level (0 dB-30 dB) Using the I-vector with 100, 200 and 400 Dimensions at UBM Mixture Size 256 for TIMIT Database Under AWGN, Street Traffic, Bus Interior and Crowd Talking NSN without Handset; the Best SIAs are found in: Table 6.5 to Table 6.8, Respectively

vectors dimensions (100 and 200) at UBM mixture size 256. Moreover, increasing the I-vector dimensions reduced the performance accuracy, while the lowest accuracy was (91.67%) for the highest I-vector dimension with 800 dimension, due to insufficient data having been trained. In addition, increasing the UBM mixture size increases the SIA and the best SIA was achieved at mixture size 256 for small I-vector dimensions such as at 100, 200 dimensions. However, in the higher dimensions of the I-vectors (400, and 800) which are achieved through the fusion techniques, the best SIA is achieved at the largest mixture size 512.

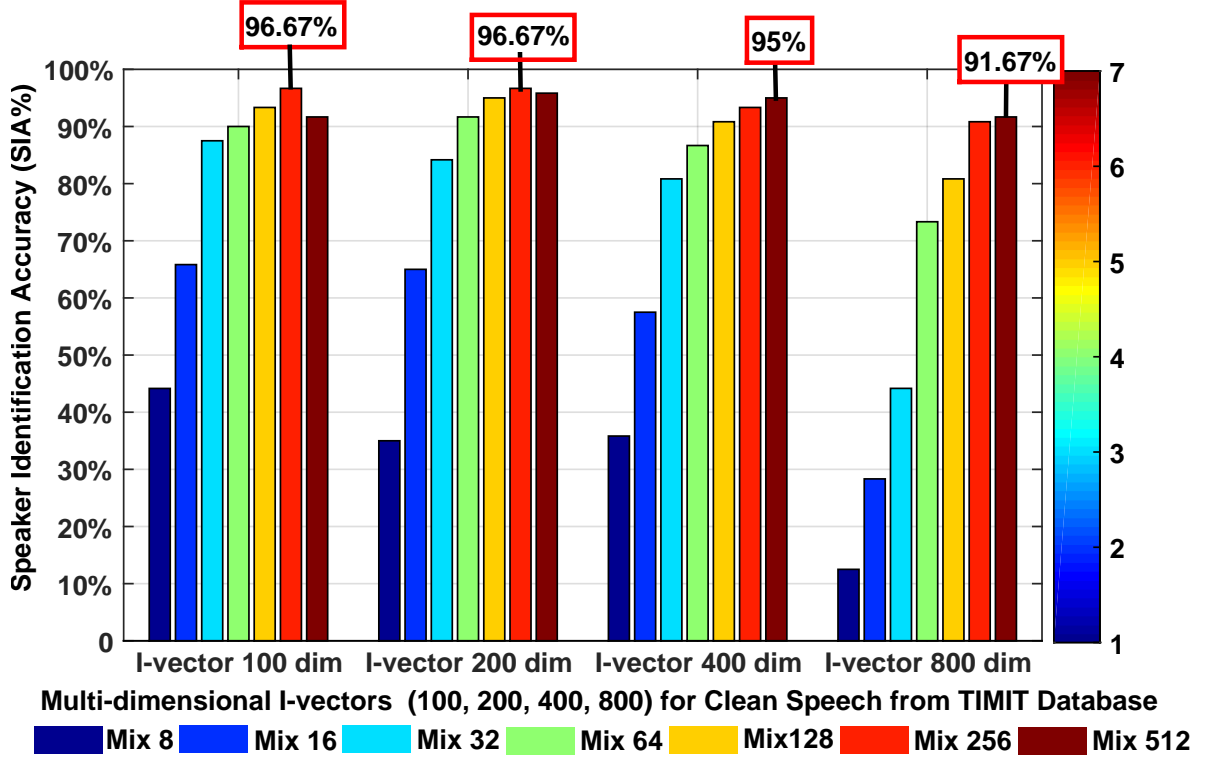


Figure 6.11: The Relationship Between SIA and Multi-Dimensional I-vectors for Different UBM Mixture Sizes for the original speech recordings of TIMIT Database

### 6.5.2 The Effects of Mixture Sizes, SNR Levels and Feature Combinations of I-vectors Without/With Fusion for SIA in TIMIT Database Evaluations

Three major simulations were performed using the TIMIT database to answer the second question  $Q_{6.2}$  stated in the discussion. These simulations produced the results in Table 6.3 to Table 6.5. All the simulations consisted of four feature combinations for I-vectors without fusion, FWMFCC, CMVNMFCC, FWPNCC and CMVNPNC, represented by the symbols  $f_1$ ,  $f_2$ ,  $g_1$ ,  $g_2$  respectively, as explained in Fig. 6.12. Furthermore, the I-vector with the highest SIA of the I-vector MFCC features (FWMFCC and CMVNMFCC) was fused with the corresponding I-vector PNCC features (FWPNCC and CMVNPNC). In addition, seven I-vector fusion methods were used based on four feature combinations with the I-vector, denoted by  $F_1$ ,  $F_2$ ,  $F_3$  for weighted sum, maximum and mean fusion respectively, as proposed in [76]. Similarly, the symbols  $F_4$ ,  $F_5$  and  $F_6$  represent the fusion for cumulative, concatenated and interleaved fusion respectively.

## 6.5 Experimental Results and Discussions

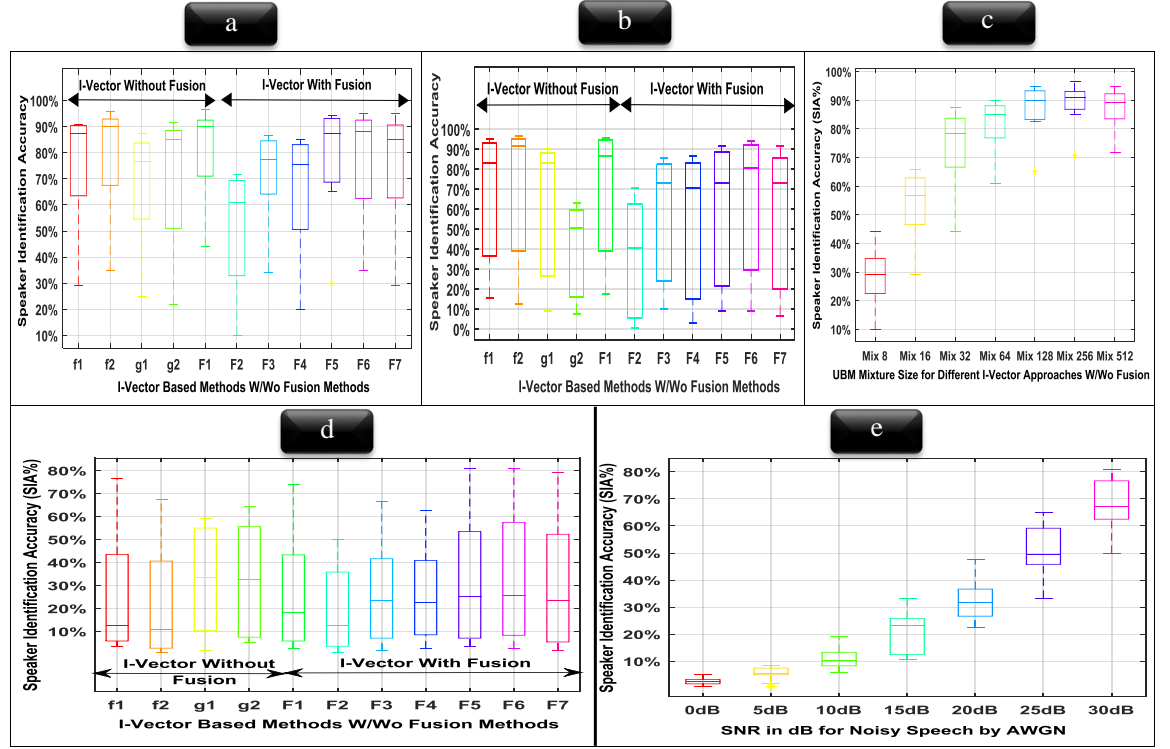


Figure 6.12: Box Plots for TIMIT Database Evaluation in original speech recordings and AWGN Noisy Speech Based on I-vector With/Without (W/WO) Fusion : Simulation 1 Represented by Part (a) and Part (c); Simulation 2 Represented by Part (b) and Simulation 3 Represented by Part (d) and Part (e): where  $\mathbf{f}_1$  and  $\mathbf{f}_2$ ,  $\mathbf{g}_1$  and  $\mathbf{g}_2$  are FWMFCC, CMCNMFCC, FWPNC and CMVNPNC I-vector Features with d-Dimension; Fusion Sets Symbols  $\mathbf{F}_1$ ,  $\mathbf{F}_2$ ,  $\mathbf{F}_3$  are d-Dimension I-vectors for Weighted Sum, Maximum and Mean Fusion.  $\mathbf{F}_4$  is d-Dimension Cumulative Fusion I-vector,  $\mathbf{F}_5$  and  $\mathbf{F}_6$  are Concatenated and Interleaving Fusion I-vectors with 2d Dimension,  $\mathbf{F}_7$  is Concatenated Fusion for the Four Feature Combinations of the I-vectors with 4d-Dimension.

However, both concatenated and interleaved fusion gave a double I-vector dimension. Finally,  $\mathbf{F}_7$  represents the fusion of all four feature combinations of the I-vector to create a new I-vector four times the original dimension.  $\dot{\mathbf{i}}$  is the I-vector for the normalized MFCC features, and had the highest SIA for CMVNMFCC and FWMFCC, denoted by  $\mathbf{f}_1$  and  $\mathbf{f}_2$ ,  $\ddot{\mathbf{i}}$  is the normalized PNCC I-vector features, which had the highest SIA for FWPNC and CMVNPNC, denoted by  $\mathbf{g}_1$  and  $\mathbf{g}_2$ . In addition,  $\mathbf{i}_{WSF}$ ,  $\mathbf{i}_{Maximum}$  and  $\mathbf{i}_{Mean}$  are the weighted sum, maximum and mean fusion I-vectors with d-dimension I-vector, denoted by  $\mathbf{F}_1$ ,  $\mathbf{F}_2$ ,  $\mathbf{F}_3$  respectively. Also,  $\mathbf{i}_{Cumulative}$ ,  $\mathbf{i}_{Concatenated}$  and  $\mathbf{i}_{interleaving}$  are Cumulative, Concatenated and Interleaved fusion I-vectors which are denoted by  $\mathbf{F}_4$ ,  $\mathbf{F}_5$  and  $\mathbf{F}_6$ , respectively.  $\mathbf{F}_4$  was with the d-dimension I-vector, while both  $\mathbf{F}_5$  and  $\mathbf{F}_6$  had 2d-dimension for the I-vector;  $\mathbf{F}_7$  was concatenated fusion of the four feature

## 6.5 Experimental Results and Discussions

---

combinations of the I-vectors with 4d-dimension. The mathematical models for all fusion methods are explained by the equations from 6.2 to 6.8 in subsection 6.3.3. This work has presented three main simulations: Simulation 1 is represented by Parts (a) and (c); Simulation 2 is represented by Part (b); and Simulation 3 is represented by Part (d) and Part (e). However, Simulation 1, represented by Part (a) and Part (c) in Fig.6.12, is based on the I-vector for original speech recordings TIMIT speech with a wide range of UBM mixture sizes. The results were taken from Table 6.3. In addition, the combination features I-vectors  $\mathbf{f}_1$ ,  $\mathbf{f}_2$ ,  $\mathbf{g}_1$ ,  $\mathbf{g}_2$  are proposed for 100 dimension without fusion and classified by ELM with 100 hidden neurons, whereas the symbols  $\mathbf{F}_1$ ,  $\mathbf{F}_2$ ,  $\mathbf{F}_3$  and  $\mathbf{F}_4$  used a 100 I-vector dimension with 100 hidden neurons for the ELM classifier. However, a double I-vector dimension (200) was created in both concatenated and interleaved fusion represented by  $\mathbf{F}_5$  and  $\mathbf{F}_6$  with 200 hidden neurons for the ELM. Moreover, the concatenated  $\mathbf{F}_5$  used two vertically concatenated features for I-vectors (features which give the best SIA for MFCC with the corresponding PNCC). Finally,  $\mathbf{F}_7$  was used for concatenation fusion of all four feature combinations of I-vectors with 400 dimension and 250 hidden neurons for the ELM, to give the best results empirically. Both Part (a) and Part (c) were taken for Simulation 1. Part (a) was mainly used to focus on the accuracy, while Part (c) emphasised the UBM mixture size. Fusion methods show improvement using a weighted sum between MFCC and PNCC features with the I-vectors, and a 1.76% improvement compared with traditional GMM-UBM. The highest SIA for the I-vector approach was 96.67%, as in Part (a) compared with 95% from GMM-UBM approach. It shows the best UBM size is 256, as explained in Part (c).

Likewise, for Simulation 2, represented in Part (b) of Fig.6.12, the results were taken from Table 6.4 and the original speech recordings of TIMIT database were evaluated for the 200 dimension for the I-vector combinations  $\mathbf{f}_1$ ,  $\mathbf{f}_2$ ,  $\mathbf{g}_1$ ,  $\mathbf{g}_2$ , and also for fusion based I-vectors  $\mathbf{F}_1$ ,  $\mathbf{F}_2$ ,  $\mathbf{F}_3$  and  $\mathbf{F}_4$  for 200 hidden neurons. Furthermore, Simulation 2 extended the I-vector dimension to involve 400 dimensions with  $\mathbf{F}_5$  and  $\mathbf{F}_6$  using empirical hidden neurons numbers as 400 and 300, respectively, compared with 350 neurons, as used in  $\mathbf{F}_7$  to create the highest I-vector dimension at 800. The fusion methods improved the system in Part (b), but this was for limited UBM mixtures below a mixture size 256. The reason for this is that an I-vector is created from a limited number of speakers, and the

system succeeded with a small I-vector size of 200 dimension, which could be extended using fusion methods.

In Simulation 3, the results were taken from Table 6.5 noisy speech and the TIMIT database was evaluated by adding AWGN without handset for a wide range of SNR levels to I-vectors based on 100, 200 and 400 dimensions with a UBM mixture size of 256, as represented in Simulation 3. Parts (d) and (e) focused on the SIA in noisy AWGN without handset and the relationship between SIA and the SNR level. All symbols proposed in Simulation 1 and Simulation 3 had the same I-vector dimensions and number of neurons, except in  $\mathbf{F}_7$  in Simulation 3, which had 300 hidden neurons. Both concatenated and interleaved fusion based I-vector improved the SIA to 80.83% compared with 79.17% from GMM-UBM, a 2.1% improvement, compared with the GMM-UBM approach at 30 dB SNR, which was a significant improvement over other SNR levels. In Part (d) Fig.6.12, the box plot illustrates that both  $\mathbf{F}_5$  and  $\mathbf{F}_6$  had higher SIA compared with other fusion methods and slightly less SIA in  $\mathbf{F}_7$ .

### 6.5.3 Comparisons of I-vector and GMM-UBM Approaches in Terms of The Speaker Identification Accuracy

This subsection answers the third question  $Q_{6.3}$  mentioned in the discussion. Fig. 6.13 shows comparisons between the two modelling approaches to produce the two speaker identification systems used in this thesis: GMM-UBM (used in Chapter 5) and the I-vector (Chapter 6). Both systems were trained as in Part A in Fig. 6.13, and tested as in Part B. Depending on the database used, there are three comparisons. In the first comparison, the results between the I-vector and GMM-UBM approaches are presented based on the TIMIT database. Secondly, there are related comparisons of the SITW database. Thirdly, the same comparisons are made depending on the NIST 2008 database. The evaluations of the comparison of the TIMIT database for the I-vector and GMM-UBM techniques included various background noise types with/without a handset: original speech recordings, AWGN Without Handset (WOH), AWGN With Handset (WH), street traffic NSN WOH and WH, bus interior NSN WOH and WH, and finally, crowd talk NSN WOH and WH. In addition, a G.712 type handset at 16 kHz was used and each simulation was achieved by employing eleven

## 6.5 Experimental Results and Discussions

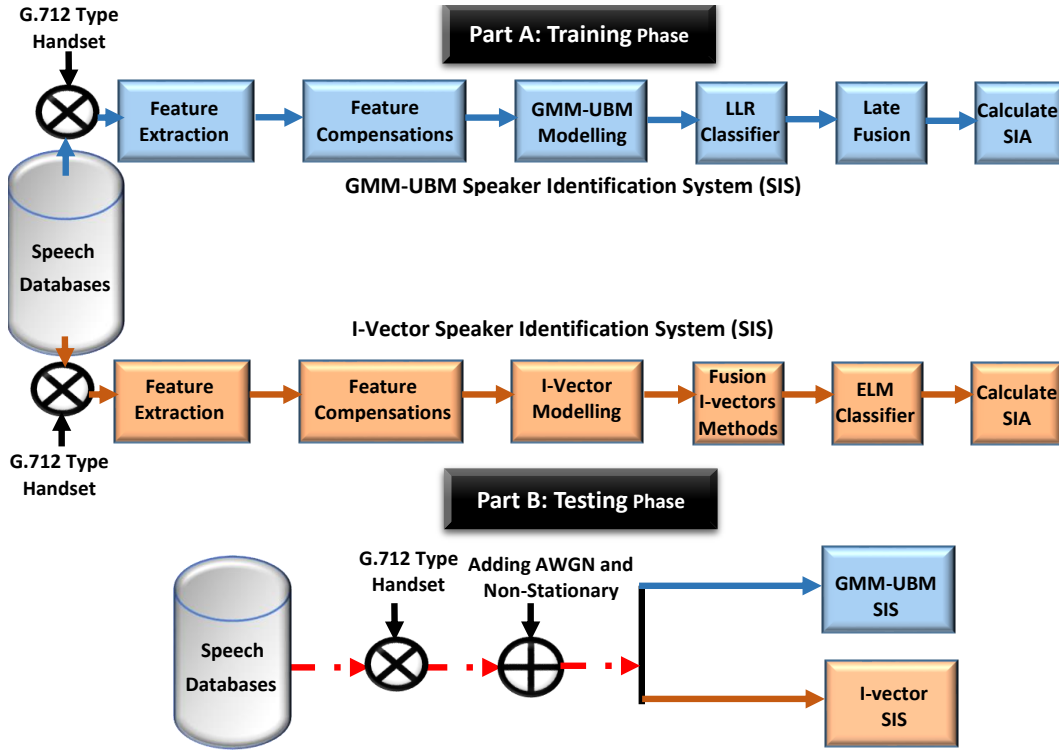


Figure 6.13: Comparison of Two Speaker Identification Frameworks Using GMM-UBM and I-vector Approaches Evaluated Under Different Environmental Conditions: Part A, Training Phase; Part B, Testing Phase

I-vectors based on four feature combinations with and without fusion methods. The feature combinations of the I-vector are: FWMFCC, CMVNMFCC, FWPNCC and CMVNPNC with a 100 I-vector dimension. There were seven other fusion methods: weighted sum, maximum, mean, cumulative I-vector fusion with d-dimension (100), concatenated and interleaved fusion with a 2d-I-vector dimension (200), and concatenated fusion with a 4d-dimension (400). In Fig. 6.14, the results are for GMM-UBM and I-vector comparisons in original speech recordings for TIMIT, and the best SIA for each mixture size was selected from both approaches regardless of feature or fusion method used. For small mixture sizes (8-64), the GMM-UBM outperformed the I-vector approach, while the I-vector showed better SIA compared with GMM-UBM at mixtures 128 and 256. However, the highest SIA was with a rate 96.67% at mixture size 256, as explained in Fig. 6.14; thereby, the mixture size 256 was used for the evaluation of all noise conditions. The results based on the I-vector approach for original speech recordings from TIMIT database was taken from Table 6.3.

Fig. 6.15 and Fig. 6.16 include the comparisons for both GMM-UBM and

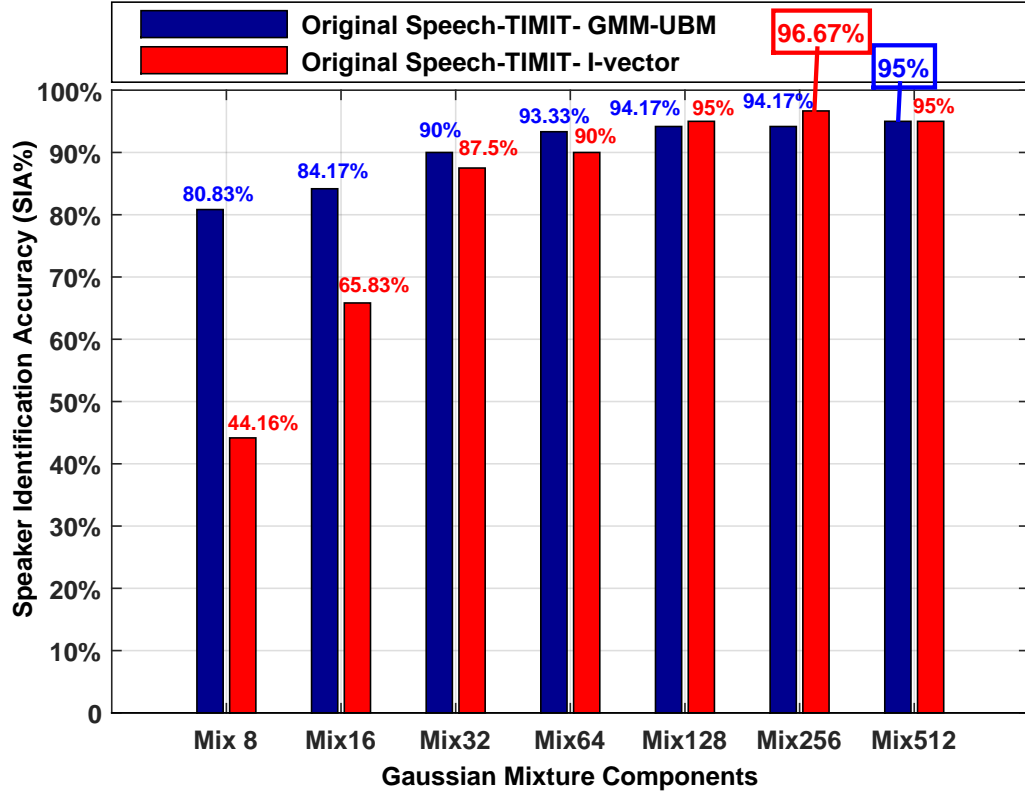


Figure 6.14: Bar Chart Plot Comparisons Between SIA Against Gaussian Mixture Components for GMM-UBM and I-vector Approaches in Terms of Original Speech Recordings From TIMIT Database

I-vector systems in AWGN, street NSN, bus NSN and crowd talking NSN with/without a handset for a wide range of SNR (0-30) dB. The results for I-vector in Fig. 6.15 were taken from Table 6.5-Table 6.8 for noisy results without handset. While other results in Fig. 6.16 were taken from Table 6.9 - Table 6.12 for noisy results with handset. The continuous coloured curves with NSN square nodes for SNR levels represent the I-vector approach, while the dash-dot coloured curves with circle nodes for SNR levels depict the GMM-UBM approach. Furthermore, the same colours were used for the same noise types for both systems. The worst performance was using the AWGN because it has a constant noise spectrum, while bus NSN achieved less reduction in SIA in the presence of noise, compared to all other non-stationary noise types.

On the other hand, both SIAs for street and crowd talking NSN were located between AWGN and the bus NSN. The relationship between the SIA for both GMM-UBM and I-vector approaches is explained in Fig. 6.15 and Fig. 6.16 with different noise conditions with/without the handset. Secondly, Fig. 6.17 shows the

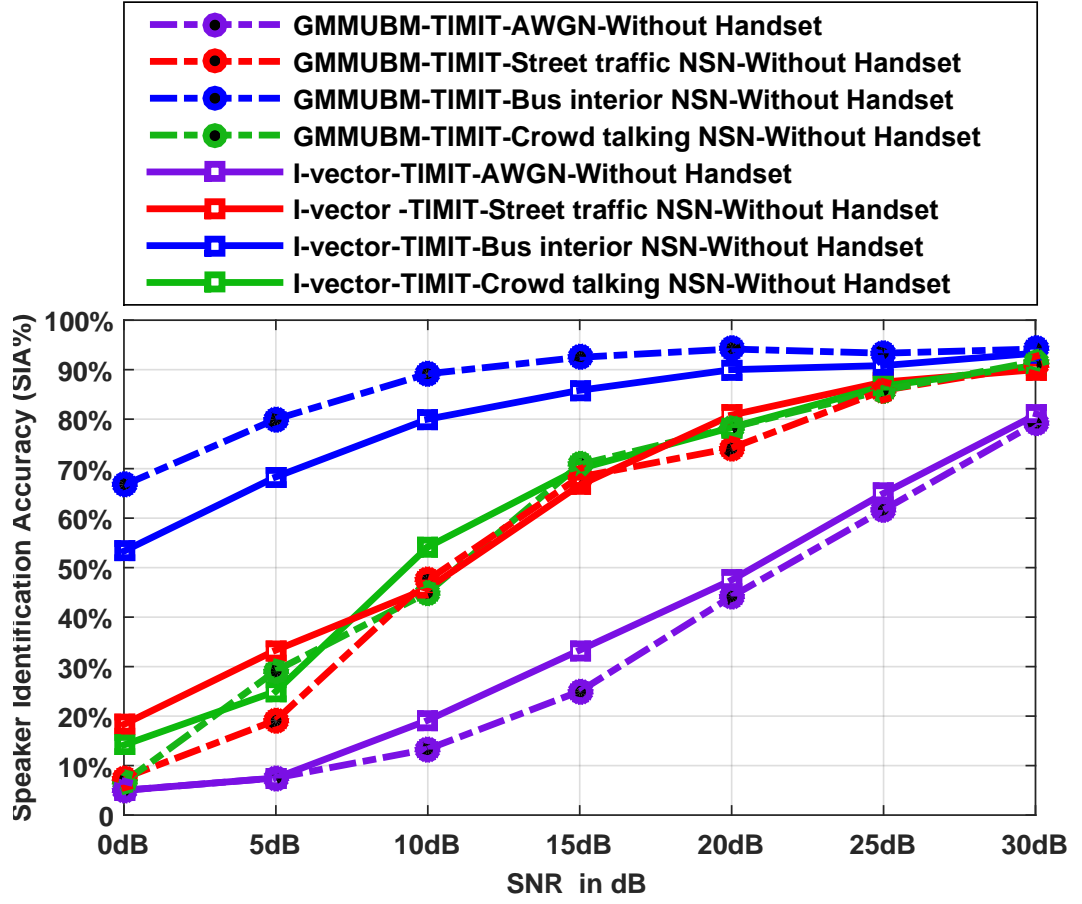


Figure 6.15: Curve Plot Comparison GMM-UBM and I-vector Approaches for AWGN and NSN without Handset at UBM Mixture Size 256 for TIMIT Database

evaluations of the SITW database for the I-vector and GMM-UBM techniques based on speech utterances from the SITW database without noise and handset. In Fig. 6.17, the best SIA for the I-vector approach was achieved using the SITW database is 85.83% at UBM mixture size 512, compared with 82.5% for the GMM-UBM with the same mixture size. However, Fig. 6.18 illustrates the noisy speech from SITW database and the best SIA values using the I-vector approach at 30 dB were 84.17%, 84.17%, 86.67% and 85% under AWGN-WH, street traffic-WH NSN, bus interior-WH NSN and crowd talking-WH NSN, respectively. However, while the best SIAs using the GMM-UBM at 30 dB are 78.33% , 81.67%, 80.83% and 82.5% for AWGN-WH, street traffic-WH NSN, bus interior-WH NSN and crowd talking-WH NSN, respectively. It is evident that the SIA for the I-vectors outperform the corresponding SIA for the GMM-UBM approach for all environments. Thirdly, Fig. 6.19 depicts the evaluations of the NIST 2008 database for the I-vector and GMM-UBM techniques based on speech utterances from NIST 2008 database without noise and handset. Fig. 6.19 shows that the highest SIA for



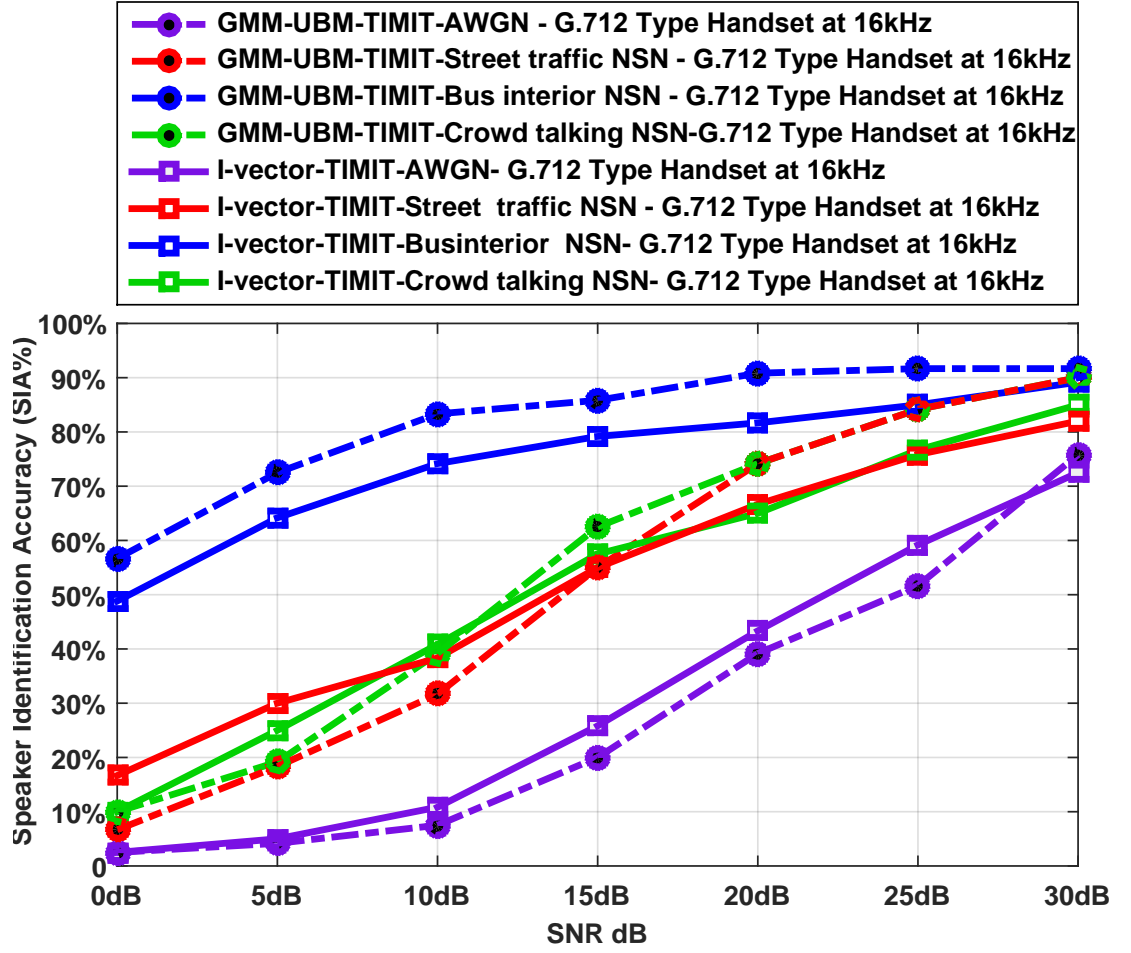


Figure 6.16: Curve Plot Comparison GMM-UBM and I-vector Approaches for AWGN and NSN with G.712 Type Handset at 16 kHz at UBM Mixture Size 256 for TIMIT Database

the I-vector approach using the NIST 2008 database was 96.67% at UBM mixture size 256 compared with 95.83% at mixture size 64 for the GMM-UBM. It is evident that the SIA is better for GMM-UBM than the corresponding SIA I-vector for the range of mixture sizes (8-64). On the other hand, the SIA was either equal or the I-vector approach outperformed the GMM-UBM for the remaining ranges of mixtures (128-512). Fig. 6.20 illustrates the noisy speech from NIST 2008 database under AWGN-WH, street-WH NSN, bus interior-WH NSN and crowd talking-WH NSN, respectively. In addition, the results were taken from Table 6.19 - Table 6.22, respectively. The best SIAs at 30 dB were 81.67% , 78.33%, 87.5% and 85% for the AWGN-WH, street traffic-WH NSN, bus interior-WH NSN and crowd talking-WH NSN, respectively. However, the best SIAs using the GMM-UBM at 30 dB were 26.67% , 80%, 92.5% and 84.17% for AWGN-WH, street traffic-WH NSN, bus interior-WH NSN and crowd talking-WH NSN, respectively. It is evident that the

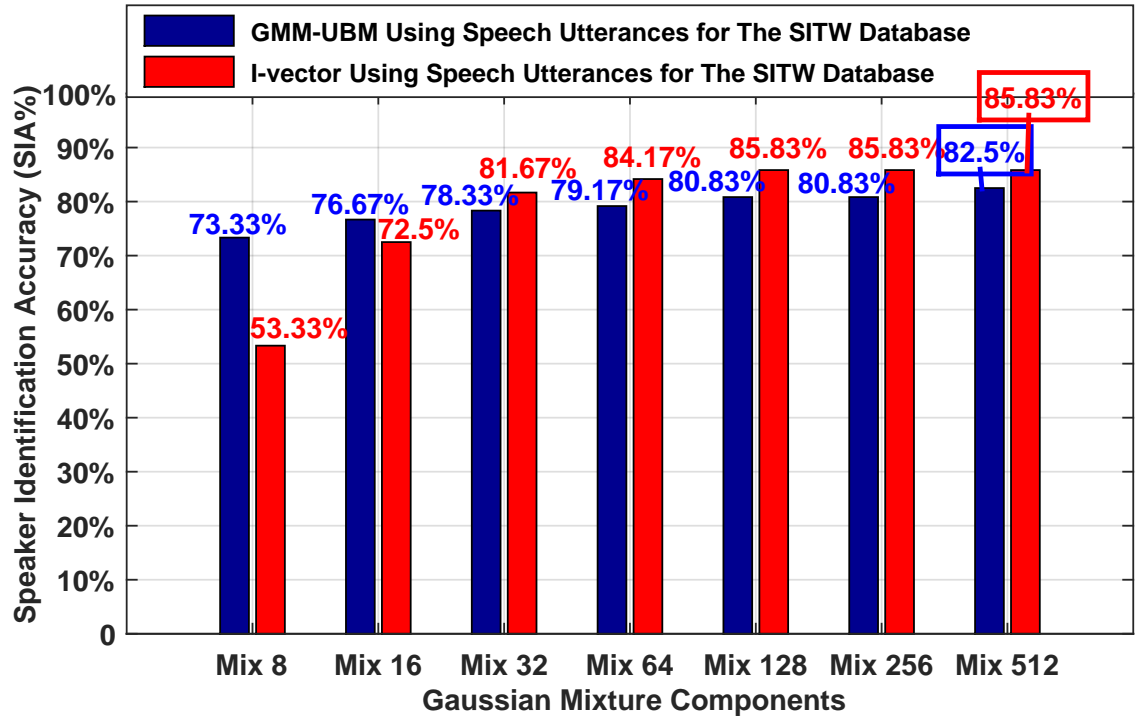


Figure 6.17: The Comparison Between the SIA for the GMM-UBM and I-vector Approaches for Speech Utterances From SITW Database without Noise and Handset

SIA for the I-vectors outperformed the corresponding SIA for the GMM-UBM approach for all mentioned environments.

## 6.6 Recent Works Related to I-vector and GMM-UBM Techniques Speaker Identification

This section summarizes the related work based on I-vector and GMM-UBM approaches, and other approaches are also considered, as explained in Table 6.2. This table presents previous work in [76] and other state of the art methods [1], [3], [46], [47], [48] and [52]. This shows the state of the art methods using the I-vector and GMM-UBM approaches, which can be compared with the results obtained in this chapter. It is evident that the current work using the I-vector outperforms other work using original speech recordings when testing with the TIMIT and NIST 2008 databases, and in other challenging environments. It is clear from the simulations and results discussed earlier in this chapter that better SIA based on the I-vector were achieved compared with the GMM-UBM under

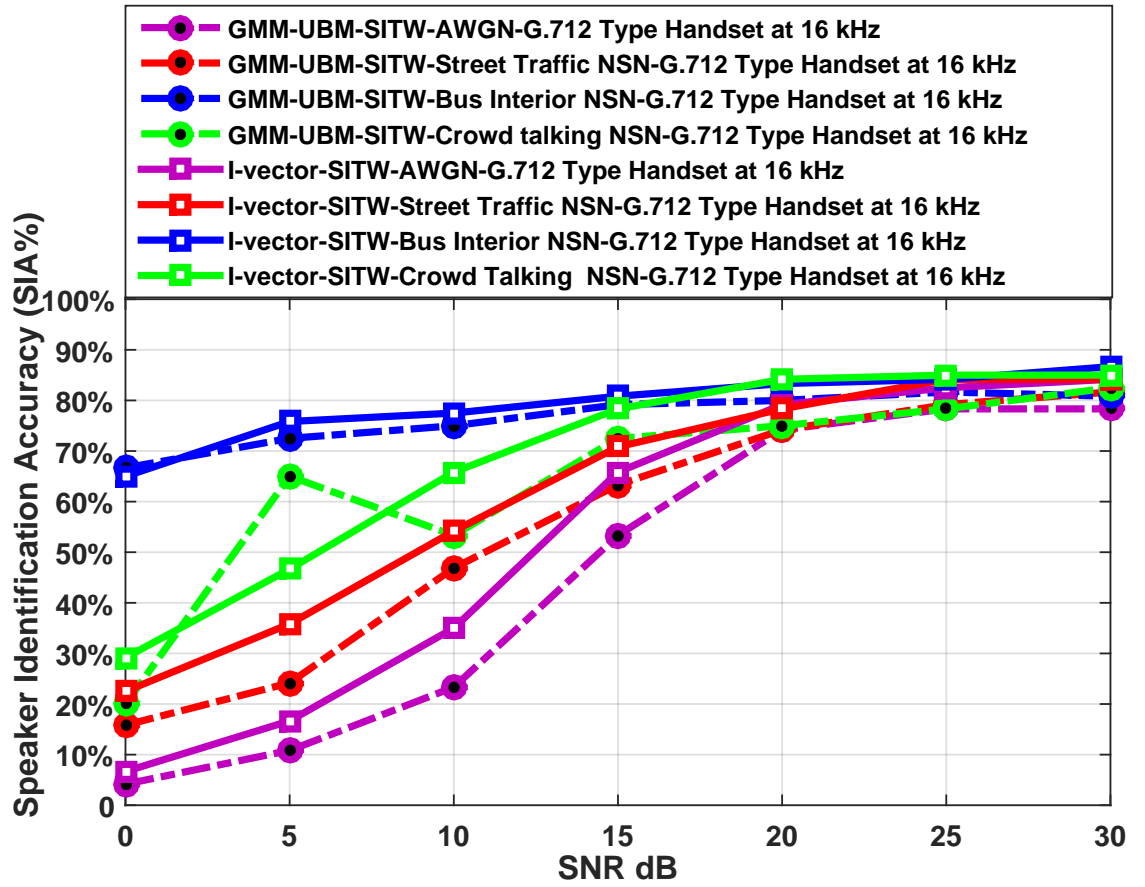


Figure 6.18: The Comparison Between the SIA for the GMM-UBM and I-vector Approaches for AWGN, Street Traffic NSN, Bus Interior NSN and Crowd Talking NSN with G.712 Type Handset at 16 kHz (at UBM mixture size 256 ) for The SITW Database

original speech recordings for TIMIT, and were equal for the NIST 2008 databases. The results also outperformed all original speech recordings measurements by other researchers. For TIMIT, the proposed I-vector approach achieved higher SIA under AWGN compared with the previous study on the GMM-UBM system, compared with other work. In contrast, the previous work in [76] with GMM-UBM had better SIA than the proposed I-vector for AWGN WH, in line with other work. In addition, for non-stationary background noise WH, the performance accuracy of GMM-UBM was better than the I-vector at SNR 30dB, but this was reversed for some SNR levels. Finally, in [52], it seems the SIA for street noise was higher than in the proposed work, but this was achieved using a different noise database with 630 speakers.

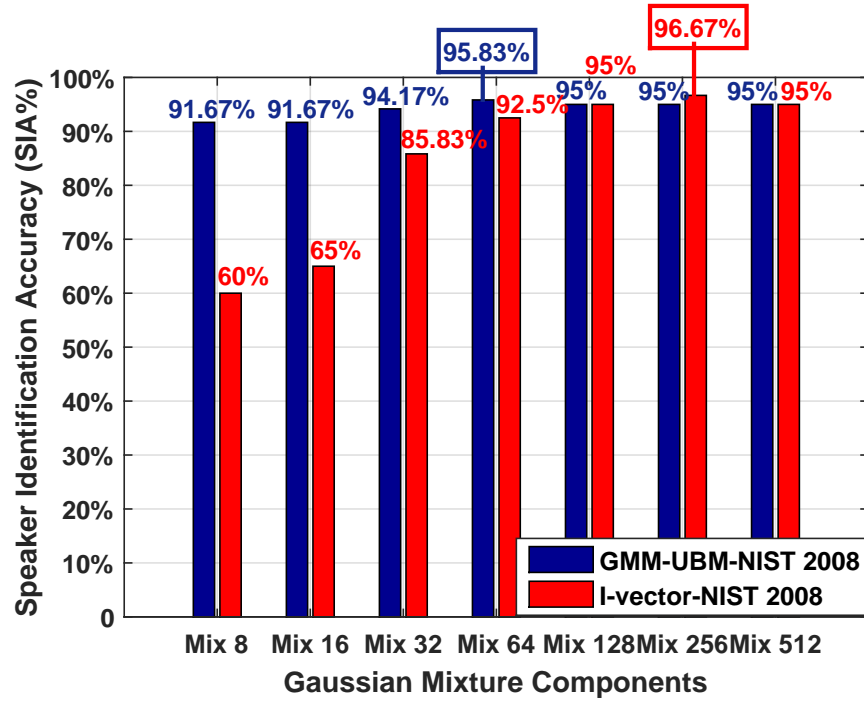


Figure 6.19: The Comparison Between the SIA for the GMM-UBM and I-vector Approaches for Speech Utterances From NIST 2008 Database without Noise and Handset

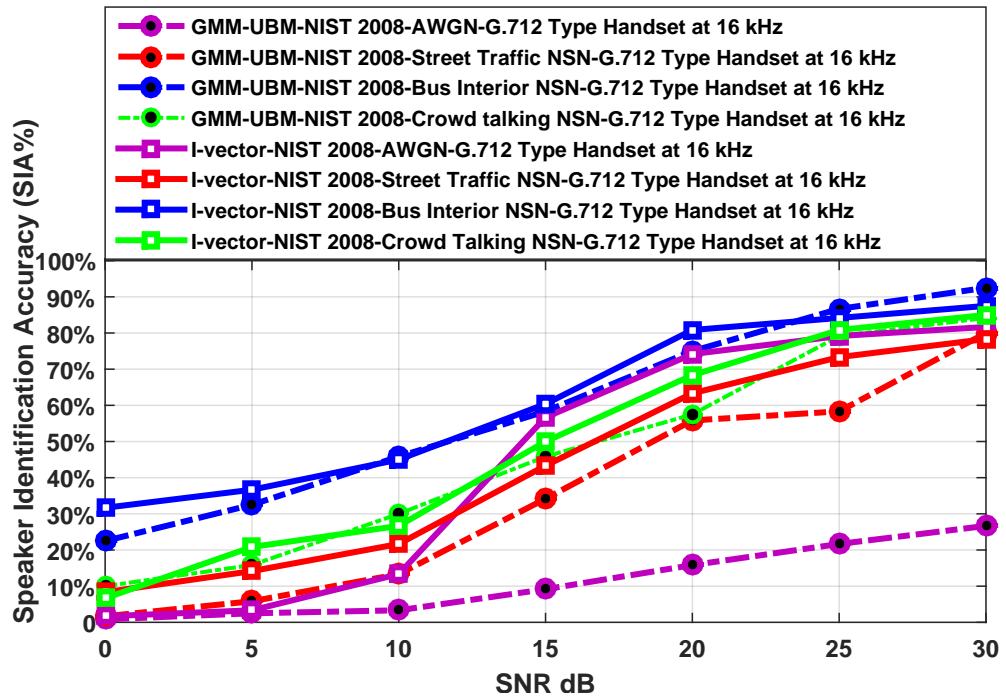


Figure 6.20: The Comparison Between the SIA for the GMM-UBM and I-vector Approaches for AWGN, Street Traffic NSN, Bus Interior NSN and Crowd Talking NSN with G.712 Type Handset at 16 kHz (at UBM Mixture Size 256 ) for the NIST 2008 Database

## 6.7 Summary

In this chapter, four feature combinations with seven fusion methods based on I-vector were investigated to develop a novel closed-set text independent speaker identification system. The new system exploited fusion based multi-dimensional I-vectors and classification with a single layer ELM neural network. The system was tested by four different databases separately (each database is tested alone): TIMIT, NTIMIT, SITW and NIST 2008 databases with 120 speakers for each database (total 480 speakers, 4,800 speech utterances). The fusion techniques were used to improve the SIA in the original speech recordings and reduce the reduction in SIA in the presence of noise and/or handset. This chapter can be summarized by the following points:

- Firstly, the identification accuracy for the I-vector seems to outperform the GMM-UBM for most environments for the SITW and NIST 2008 databases and this is due to using the combination of the ELM and the I-vector which improves the SIA. However, with the TIMIT database the system outperformed the GMM-UBM techniques for original speech recordings, and also outperformed under AWGN WOH, so that it seems better for some SNR levels with street and crowd talking. In contrast, for bus interior NSN, the GMM-UBM achieved less reduction in SIA compared with the I-vector approach. Additionally, fusion techniques may mitigate the reduction in SIA caused by different noise environments and the handset effect, whereas weighted fusion generally seem to be the best of all the feature and fusion methods used. However, the new database using SITW demonstrated that the identification accuracy achieved by the I-vector approach is better than the corresponding results for the GMM-UBM method for all challenging environments. With the NIST 2008 database, it seems the output from GMM-UBM is better than small mixture sizes I-vector, but this was reversed with an increase in the UBM mixture size, to give slightly better results in the original speech recordings. The I-vector gives slightly higher SIA than GMM-UBM in different noise types, except for the bus NSN, where the GMM-UBM outperforms the I-vector. In the TIMIT database, the I-vector approach has better performance than GMM-UBM in original speech recordings and for AWGN without handset, while in other types of NSN the

I-vector outperforms the GMM-UBM in some SNR levels.

- Secondly, this chapter considered a simple, efficient ELM classifier for the speaker identification task.
- Thirdly, the smallest I-vector with 100 and 200 dimensions has higher SIA compared with other I-vector dimensions of 400 and 800.
- Fourthly, the best UBM mixture size is 256, while the best SIA in noise and handset conditions was achieved at SNR 30 dB for all databases used. Fifthly, in noisy conditions the worst SIA is achieved at AWGN due to the stationary spectrum for the noise, while the highest SIA is obtained with bus interior NSN. In addition, the SIA for the street and crowd talking NSN were between the SIA values for the bus NSN and AWGN.
- Finally, the best fusion method according to the highest SIA for the TIMIT database was the weighted sum fusion, while in the SITW and NIST 2008 databases the best fusion type was concatenated fusion with 2d I-vector dimension. In addition, some other fusion types were also useful to achieve improvements in SIA for different environments and in different SNR level, as explained clearly in the tables of results in the appendix section for this chapter.

The chapter also answered three important questions and they are: the first question  $Q_{6.1}$  concerns how far the I-vector dimensions affect the SIA; the second question  $Q_{6.2}$  is how far the UBM mixture sizes, SNR level, feature combination of I-vector, with and without fusion, affected the SIA; the final question  $Q_{6.3}$  is which is best, GMM-UBM or I-vector approaches for speaker identification. The next chapter will consider the thesis conclusion and the future work related to the speaker identification system.

Table 6.2: Related Work with I-vector and GMM-UBM Proposed Work

Recent works Related to I-vector and GMM-UBM Techniques Speaker Identification						
Approaches	Database	Number of Speakers	The best feature/ fusion Based	Environments	The best SIA	Authors
Fusion based GMM-UBM	TIMIT	120 speakers	Weighted sum	Clean	95%	Al-kaltakchi et al. [76] [2016]
Fusion based GMM-UBM	TIMIT	120 speakers	Maximum fusion	AWGN	79.17%(30dB)	Al-kaltakchi et al. [76] [2016]
Fusion based GMM-UBM	TIMIT	120 speakers	FWMFCC-feature	AWGN with Handset	75.83%(30dB)	Al-kaltakchi et al. [76] [2016]
I-vector Approach	NIST-2008	400 registered speakers	Without fusion	Clean	49.5%	[48] [2014]
I-vector Approach	NIST-2008	400 speakers	Without fusion	White noise	39.3%(15dB)	[48] [2014]
GMM-UBM Approach	NIST-2008	400 speakers	Without fusion	Clean	39.7%	[48] [2014]
GMM-UBM Approach	NIST-2008	400 speakers	Without fusion	White noise	24.6%(15dB)	[48] [2014]
GMM-UBM-ZT norm	NIST-2008	400 speakers	Without fusion	Clean	42.5%	[48] [2014]
GMM-UBM-ZT norm	NIST-2008	400 speakers	Without fusion	White noise at SNR 15 dB	29.7% (15dB)	[48] [2014]
I-vector+LDA+WCCN	Corpus designed and MIT mobile phone	50 speakers	Without fusion	Clean	94.14%(CDS)	[46] [2014]
I-vector+LDA+WCCN		50 speakers	Without fusion	Clean	92.36% (SVM)	[46] [2014]
I-vector 400 Dim	YouTube	1,000 speakers	Without fusion	Clean	92% test (10s)	[47] [2014]
LDA 300 Dim		1,000 speakers	Without fusion	Clean	96.1% (20s)	[47] [2014]
Fusion Based GMM	TIMIT	120 speakers	Weighted sum	Clean	93.88%	[1] [2012]
GMM-UBM without fusion	TIMIT	64 speakers	Without fusion	Clean	94.5%	[3] [2011]
				AWGN with Handset	74.2%(30dB)	[3] [2011]
New model with GMM	TIMIT	630 speakers	Without fusion	Clean	96.51%	[52] [2007]
The best result at Mix 128				Street NSN (20dB)	92.86%	[52] [2007]

## 6.8 Appendix 6.1

Table 6.3: Simulation 1: The Speaker Identification Accuracy (SIA) as a Function of the UBM Mixture Sizes  $\{8, 16, 32, 64, 128, 256, 512\}$  for the I-vector Approach for 100, 200 and 400 Dimensions for the Original Speech Recordings of TIMIT Database

Simulation 1: The SIA for Original Speech Recordings of <b>TIMIT Database</b>							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
Feature based I-vector							
Without Fusion							
With 100 dimension							
$NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	29.17%	55.83%	87.5%	86.6%	90.83%	90.83%	89.17%
CMVNMFCC( $\mathbf{f}_2$ )	35%	61.67%	85%	90%	93.33%	95.83%	91.67%
FWPNCC ( $\mathbf{g}_1$ )	25%	49.17%	70.83%	76.67%	82.5%	87.5%	84.17%
CMVNPNC ( $\mathbf{g}_2$ )	21.67%	45.83%	66.67%	85%	86.67%	91.67%	89.17%
Fusion Decision	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$
Fusion based I-vector							
With 100 dimension							
$NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	44.16%	65.83%	85%	88.33%	93.33%	<b>96.67%</b>	89.17%
$i_{WSF}$ at $\omega_2=0.8$	40%	65.83%	85%	89.17%	93.33%	93.33%	90%
$i_{WSF}$ at $\omega_3=0.77$	43.33%	65.83%	86.67%	90%	90.83%	93.33%	89.17%
$i_{WSF}$ at $\omega_4=0.7$	35%	65%	80.83%	86.67%	91.67%	91.67%	86.67%
$i_{Maximum}$	10%	29.17%	44.17%	60.83%	65%	70.83%	71.67%
$i_{Mean}$	34.17%	63.33%	66.67%	77.5%	85.83%	86.67%	80.83%
$i_{Cumulative}$	20%	45.83%	65%	75.33%	82.5%	85%	83.33%
Fusion based I-vector							
With 200 dimension							
$i_{Concatenated} (2d)$							
$NoHN = 100$	27.5%	55%	68.33%	81.67%	84.16%	85%	87.5%
$NoHN = 150$	27.5%	56.67%	81.67%	86.67%	90%	95%	91.67%
$NoHN = 200$	30%	65%	80%	87.5%	94.17%	93.33%	92.5%
$i_{interleaving} (2d)$							
$NoHN = 100$	27.5%	56.67%	70.83%	78.33%	85%	87.5%	85.83%
$NoHN = 150$	34.17%	61.67%	84.17%	90%	95%	92.5%	93.33%
$NoHN = 200$	35%	56.67%	80%	88.33%	95%	92.5%	92.5%
Fusion based I-vector							
With 400 dimension							
$i_{Concatenated} (4d)$							
$NoHN = 100$	27.5%	39.17%	68.33%	70.83%	80.83%	85%	81.66%
$NoHN = 150$	35.83%	49.17%	80.83%	84.17%	87.5%	90.83%	87.5%
$NoHN = 200$	32.5%	55.83%	78.33%	84.17%	90.83%	90.83%	93.33%
$NoHN = 250$	29.17%	57.5%	78.33%	85%	90%	90.83%	95%
$NoHN = 300$	30%	45%	73.33%	86.67%	88.33%	93.33%	90%



## 6.8 Appendix 6.1

Table 6.4: Simulation 2: The Speaker Identification Accuracy (SIA) as a Function of the UBM Mixture Sizes {8, 16, 32, 64, 128, 256, 512 } for the I-vector Approach for 200, 400 and 800 Dimensions for Original Speech Recordings for the TIMIT Database

Simulation 2: The SIA for Original Speech Recordings for the <b>TIMIT Database</b>							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
Feature based I-vector							
Without Fusion							
With 200 dimension							
$NoHN = 200$							
FWMFCC ( $\mathbf{f}_1$ )	15.83%	29.17%	58.33%	83.33%	93.33%	95%	92.5%
CMVNMFCC ( $\mathbf{f}_2$ )	12.5%	32.5%	58.33%	91.67%	94.17%	<b>96.67%</b>	95.83%
FWPNCC ( $\mathbf{g}_1$ )	9.16%	19.17%	50%	90%	83.33%	87.5%	88.33%
CMVNPNC ( $\mathbf{g}_2$ )	7.5%	12.5%	26.67%	50.83%	55.83%	63.33%	60.83%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$
Fusion based I-vector							
With 200 dimension							
$NoHN = 200$							
$i_{WSF}$ at $\omega_1=0.9$	15%	30.83%	65%	85.83%	95%	95.83%	94.17%
$i_{WSF}$ at $\omega_2=0.8$	15%	24.17%	64.17%	86.67%	94.17%	95%	94.17%
$i_{WSF}$ at $\omega_3=0.77$	12.5%	30%	58.33%	86.67%	92.5%	95%	93.33%
$i_{WSF}$ at $\omega_4=0.7$	17.5%	26.67%	55.83%	82.5%	91.67%	93.33%	92.5%
$i_{Maximum}$	0.83%	3.33%	13.33%	40.83%	52.5%	65.83%	70.83%
$i_{Mean}$	10%	17.5%	43.33%	73.33%	77.17%	84.17%	85.83%
$i_{Cumulative}$	3.33%	12.5%	23.33%	70.83%	78.33%	85%	86.67%
Fusion based I-vector							
With 400 dimension							
$i_{Concatenated}$ (2d)							
$NoHN = 200$	8.33%	26.67%	48.33%	75%	76.67%	90%	92.5%
$NoHN = 250$	10.83%	20.83%	55.83%	78.33%	80.83%	86.67%	90%
$NoHN = 300$	12.5%	24.17%	57.5%	81.67%	84.17%	91.67%	90.83%
$NoHN = 400$	9.17%	14.17%	43.33%	73.33%	82.5%	90.83%	91.67%
$i_{interleaving}$ (2d)							
$NoHN = 200$	10.83%	25.83%	47.5%	75%	79.17%	89.17%	92.5%
$NoHN = 250$	10%	27.5%	56.67%	80%	84.17%	94.17%	91.67%
$NoHN = 300$	9.17%	22.5%	50.83%	80.83%	88.33%	94.17%	93.33%
$NoHN = 400$	6.67%	9.17%	35.83%	73.33%	84.17%	92.5%	92.5%
Fusion based I-vector							
With 800 dimension							
$i_{Concatenated}$ (4d)							
$NoHN = 350$	6.67%	15%	35.83%	73.33%	80.83%	87.5%	91.67%
$NoHN = 300$	10.83%	17.5%	44.17%	72.5%	80.83%	90.83%	89.17%
$NoHN = 250$	12.5%	28.33%	40.83%	71.67%	79.17%	88.33%	90.83%

## 6.8 Appendix 6.1

Table 6.5: Simulation 3: The SIA for the I-vector Approach for 100, 200 and 400 Dimensions Under AWGN at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB Without Handset at UBM Mixture Size 256 for the TIMIT Database

Simulation 3: The SIA for AWGN Without Handset for The <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector							
Without Fusion							
With 100 dimension							
$NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	3.33%	5%	8.33%	12.5%	26.67%	49.17%	76.67%
CMVNMFCC( $\mathbf{f}_2$ )	0.83%	1.67%	5.83%	10.83%	27.5%	45%	67.5%
FWPNCC ( $\mathbf{g}_1$ )	1.67%	7.5%	19.17%	33.33%	45%	59.17%	58.33%
CMVNPNCN ( $\mathbf{g}_2$ )	5%	5.83%	12.5%	32.5%	47.5%	58.33%	64.17%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fusion based I-vector							
With 100 dimension							
$NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	1.67%	5%	8.33%	12.5%	23.33%	46.67%	69.17%
$i_{WSF}$ at $\omega_2=0.8$	1.67%	3.33%	7.5%	16.67%	20.83%	46.67%	68.33%
$i_{WSF}$ at $\omega_3=0.77$	1.67%	3.33%	7.5%	16.67%	21.67%	49.17%	74.17%
$i_{WSF}$ at $\omega_4=0.7$	2.5%	4.17%	7.5%	18.33%	20%	50%	68.33%
$i_{Maximum}$	1.67%	0.83%	9.16%	12.5%	36.67%	33.33%	50%
$i_{Mean}$	1.67%	8.33%	6.66%	23.33%	29.17%	45.83%	66.67%
$i_{Cumulative}$	2.5%	7.5%	11.66%	23.33%	22.5%	46.67%	62.5%
Fusion based I-vector							
With 200 dimension							
$i_{Concatenated}$ (2d)							
$NoHN = 200$	3.33%	5%	13.33%	25%	34.17%	60%	<b>80.83%</b>
$i_{interleaving}$ (2d)							
$NoHN = 200$	2.5%	6.67%	13.33%	25.83%	35%	65%	<b>80.83%</b>
Fusion based I-vector							
With 400 dimension							
$i_{Concatenated}$ (4d)							
$NoHN = 200$	0.83%	3.33%	5.83%	16.67%	29.17%	45.83%	65.83%
$NoHN = 300$	1.67%	4.17%	9.17%	23.33%	39.17%	56.67%	79.17%
$NoHN = 400$	1.67%	0.83%	9.17%	20.83%	35.83%	52.5%	74.17%

## 6.8 Appendix 6.1

Table 6.6: Simulation 4: The SIA for Different Gaussian Mixture Components (GMCs) for the I-vector Approach for 100, 200 and 400 Dimensions Under Street Traffic NSN Without Handset at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the TIMIT Database

Simulation 4: The SIA for Street Traffic NSN Without Handset for the <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	13.33%	33.33%	41.67%	61.67%	77.5%	81.67%	87.5%
CMVNMFCC( $\mathbf{f}_2$ )	10%	23.33%	40%	55.83%	66.67%	75%	78.33%
FWPNCC ( $\mathbf{g}_1$ )	5.83%	15%	30%	40%	55%	60%	65.83%
CMVNPNC ( $\mathbf{g}_2$ )	5%	9.17%	29.17%	48.33%	52.5%	68.33%	71.67%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	13.33%	30%	43.33%	62.5%	80%	83.33%	<b>90%</b>
$i_{WSF}$ at $\omega_2=0.8$	13.33%	28.33%	45.83%	62.5%	80.83%	83.33%	83.33%
$i_{WSF}$ at $\omega_3=0.77$	18.33%	27.5%	44.17%	59.17%	80.33%	81.67%	85.83%
$i_{WSF}$ at $\omega_4=0.7$	15%	25.83%	42.5%	58.33%	74.17%	80%	80%
$i_{Maximum}$	4.17%	13.33%	15.83%	34.17%	46.67%	50.83%	56.67%
$i_{Mean}$	6.67%	20%	30.83%	53.33%	61.67%	71.67%	76.67%
$i_{Cumulative}$	6.67%	23.33%	31.67%	51.67%	63.33%	70%	75%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	8.33%	25%	43.33%	66.67%	75.83%	87.5%	89.17%
$i_{interleaving}$ (2d) $NoHN = 200$	7.5%	20.83%	38.33%	60.83%	75.83%	85.83%	89.17%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	5%	20.83%	42.5%	60.83%	75%	79.17%	83.33%
$NoHN = 300$	6.67%	21.67%	32.5%	56.67%	69.17%	75.83%	87.5%
$NoHN = 400$	3.33%	14.17%	30%	44.17%	64.17%	79.17%	80.83%

## 6.8 Appendix 6.1

Table 6.7: Simulation 5: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Bus Interior NSN Without Handset at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the TIMIT Database

Simulation 5: The SIA for Bus Interior NSN Without Handset for the <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	53.33%	65%	78.33%	85.83%	86.67%	90.83%	92.5%
CMVNMFCC( $\mathbf{f}_2$ )	50.83%	63.33%	70.83%	80.83%	85%	85.83%	90.83%
FWPNCC ( $\mathbf{g}_1$ )	19.17%	36.67%	51.67%	65.83%	65%	70%	67.5%
CMVNPNC ( $\mathbf{g}_2$ )	23.33%	31.67%	51.67%	63.33%	69.17%	68.33%	75%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	51.67%	67.5%	77.5%	83.33%	88.33%	89.17%	89.17%
$i_{WSF}$ at $\omega_2=0.8$	47.5%	63.33%	75%	80%	87.5%	89.17%	91.67%
$i_{WSF}$ at $\omega_3=0.77$	42.5%	64.17%	73.33%	80.83%	88.33%	88.33%	88.33%
$i_{WSF}$ at $\omega_4=0.7$	39.17%	52.5%	74.17%	81.67%	84.17%	85.83%	85.83%
$i_{Maximum}$	18.33%	43.33%	48.33%	52.5%	52.5%	58.33%	63.33%
$i_{Mean}$	34.17%	49.17%	69.17%	68.33%	72.5%	76.67%	79.17%
$i_{Cumulative}$	30%	42.5%	66.67%	70.83%	71.67%	75%	79.17%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	35%	68.33%	78.33%	82.5%	87.5%	90%	90%
$i_{interleaving}$ (2d) $NoHN = 200$	38.33%	65%	80%	83.33%	88.33%	90%	<b>93.33%</b>
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	35%	61.67%	74.17%	83.33%	90%	87.5%	87.5%
$NoHN = 300$	32.5%	65%	70.83%	79.17%	83.33%	85%	89.17%
$NoHN = 400$	21.67%	50.83%	62.5%	70.83%	81.67%	78.33%	84.17%

## 6.8 Appendix 6.1

Table 6.8: Simulation 6: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Crowd Talking NSN Without Handset at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the TIMIT Database

Simulation 6: The SIA for Crowd Talking NSN Without Handset for the <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	11.67%	23.33%	47.5%	64.17%	74.17%	84.17%	85.83%
CMVNMFCC( $\mathbf{f}_2$ )	4.17%	20.83%	36.67%	52.5%	69.17%	75%	84.17%
FWPNCC ( $\mathbf{g}_1$ )	4.17%	12.5%	23.33%	34.17%	55.83%	58.33%	67.5%
CMVNPNC ( $\mathbf{g}_2$ )	1.67%	10.83%	25.83%	43.33%	53.33%	63.33%	65.83%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	13.33%	23.33%	46.67%	61.67%	71.67%	85%	84.17%
$i_{WSF}$ at $\omega_2=0.8$	11.67%	22.5%	45.83%	61.67%	74.17%	80%	86.67%
$i_{WSF}$ at $\omega_3=0.77$	14.17%	25%	43.33%	61.67%	73.33%	76.67%	88.33%
$i_{WSF}$ at $\omega_4=0.7$	13.33%	25%	41.67%	60%	67.5%	75.83%	84.17%
$i_{Maximum}$	5.83%	10.83%	24.17%	38.33%	39.17%	54.17%	54.17%
$i_{Mean}$	5%	16.67%	33.33%	51.67%	55.83%	68.33%	74.16%
$i_{Cumulative}$	9.17%	15%	39.17%	51.67%	56.67%	67.5%	73.33%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	9.17%	19.17%	50%	70%	78.33%	86.67%	<b>90.83%</b>
$i_{interleaving}$ (2d) $NoHN = 200$	12.5%	21.67%	54.17%	69.17%	74.17%	85%	88.33%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	5.83%	24.17%	37.5%	65%	75%	79.17%	83.33%
$NoHN = 300$	8.33%	20.83%	34.17%	66.67%	72.5%	81.67%	85.83%
$NoHN = 400$	4.17%	15%	29.17%	51.67%	67.5%	72.5%	76.67%

## 6.8 Appendix 6.1

Table 6.9: Simulation 7: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under AWGN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the TIMIT Database

Simulation 7: The SIA for AWGN-WH at 16 kHz for the <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	2.5%	3.33%	5%	10.83%	30%	54.17%	<b>72.5%</b>
CMVNMFCC( $\mathbf{f}_2$ )	0.83%	2.5%	2.5%	9.17%	19.17%	45.83%	65%
FWPNCC ( $\mathbf{g}_1$ )	1.67%	3.33%	10.83%	23.33%	35.83%	55%	56.67%
CMVNPNC ( $\mathbf{g}_2$ )	0.83%	5%	7.5%	25.83%	43.33%	59.17%	56.67%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	1.67%	3.33%	5.83%	10%	28.33%	54.16%	70%
$i_{WSF}$ at $\omega_2=0.8$	1.67%	4.17%	4.17%	13.33%	30%	49.17%	70%
$i_{WSF}$ at $\omega_3=0.77$	0.83%	2.5%	3.33%	12.5%	30%	48.33%	69.17%
$i_{WSF}$ at $\omega_4=0.7$	1.67%	5%	4.17%	12.5%	25.83%	50%	69.17%
$i_{Maximum}$	0%	1.67%	4.17%	7.5%	15.83%	35%	46.67%
$i_{Mean}$	5%	16.67%	33.33%	51.67%	55.83%	68.33%	74.16%
$i_{Cumulative}$	1.67%	4.17%	5%	13.33%	32.5%	47.5%	65%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	0.83%	3.33%	9.17%	25%	41.67%	57.5%	70%
$i_{interleaving}$ (2d) $NoHN = 200$	0.83%	2.5%	7.5%	21.67%	37.5%	57.5%	70%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	0%	0.83%	9.17%	19.17%	40%	51.67%	71.67%
$NoHN = 300$	0%	0.83%	5%	16.67%	33.33%	51.67%	70%
$NoHN = 400$	0.83%	0%	4.17%	4.17%	18.33%	36.67%	57.5%

## 6.8 Appendix 6.1

Table 6.10: Simulation 8: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Street Traffic NSN with G.712 Type Handset at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the TIMIT Database

Simulation 8: The SIA for Street Traffic NSN-WH <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion With 100 dimension $NoHN = 100$ FWMFCC ( $\mathbf{f}_1$ )	15.83%	30%	37.5%	50.83%	63.33%	75.83%	80%
CMVNMFCC( $\mathbf{f}_2$ )	10.83%	21.67%	38.33%	47.5%	59.17%	71.67%	72.5%
FWPNCC ( $\mathbf{g}_1$ )	4.17%	5.83%	13.33%	29.17%	39.17%	54.17%	53.33%
CMVNPNC ( $\mathbf{g}_2$ )	4.17%	5%	12.5%	22.5%	38.33%	46.67%	64.17%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$ $i_{WSF}$ at $\omega_1=0.9$	16.67%	25.83%	35%	54.17%	66.67%	75.83%	<b>82%</b>
$i_{WSF}$ at $\omega_2=0.8$	15%	25.83%	31.67%	55%	66.67%	73.33%	80.83%
$i_{WSF}$ at $\omega_3=0.77$	11.66%	23.33%	30%	53.33%	66.67%	74.17%	80.83%
$i_{WSF}$ at $\omega_4=0.7$	15%	21.67%	29.17%	44.17%	63.33%	75%	79.17%
$i_{Maximum}$	0.83%	8.33%	8.33%	20%	31.67%	41.67%	49.17%
$i_{Mean}$	5%	13.33%	17.5%	35.83%	48.33%	65%	63.33%
$i_{Cumulative}$	5%	13.33%	15%	35.83%	43.33%	60.83%	66.67%
Fusion based I-vector With 200 dimension $i_{Concatenated}$ (2d) $NoHN = 200$	9.17%	14.17%	32.5%	48.33%	65%	75%	77.5%
$i_{interleaving}$ (2d) $NoHN = 200$	10%	14.17%	30%	49.17%	61.67%	73.33%	79.17%
Fusion based I-vector With 400 dimension $i_{Concatenated}$ (4d) $NoHN = 200$	6.67%	17.5%	26.67%	34.17%	41.67%	74.17%	70%
$NoHN = 300$	6.67%	13.33%	31.67%	34.17%	57.5%	62.5%	77.5%
$NoHN = 400$	5%	11.67%	18.33%	26.67%	43.33%	55%	67.5%

## 6.8 Appendix 6.1

Table 6.11: Simulation 9: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Bus Interior NSN With G.712 Type Handset at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the TIMIT Database

Simulation 9: The SIA for Bus Interior NSN-WH for the <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	44.17%	62.5%	74.17%	75.83%	78.33%	83.33%	<b>89.17%</b>
CMVNMFCC( $\mathbf{f}_2$ )	45.83%	53.33%	67.5%	79.17%	76.67%	83.33%	88.33%
FWPNCC ( $\mathbf{g}_1$ )	15%	23.33%	39.17%	51.67%	55.83%	61.67%	60.83%
CMVNPNC ( $\mathbf{g}_2$ )	14.17%	29.17%	35%	49.17%	57.5%	65%	67.5%
Fusion Decision	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	47.5%	64.17%	70%	79.17%	78.33%	85%	88.33%
$i_{WSF}$ at $\omega_2=0.8$	48.33%	57.5%	72.5%	75%	79.17%	79.17%	<b>89.17%</b>
$i_{WSF}$ at $\omega_3=0.77$	45%	57.5%	69.17%	71.67%	81.67%	81.67%	<b>89.17%</b>
$i_{WSF}$ at $\omega_4=0.7$	46.67%	55.83%	65.83%	72.5%	75%	82.5%	84.17%
$i_{Maximum}$	25.83%	23.33%	25%	36.67%	37.5%	50.83%	54.17%
$i_{Mean}$	32.5%	38.33%	46.67%	60.83%	60%	68.33%	68.33%
$i_{Cumulative}$	30.83%	34.17%	45.83%	58.33%	59.17%	65%	68.33%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	30%	44.17%	70.83%	71.67%	71.67%	69.17%	82.5%
$i_{interleaving}$ (2d) $NoHN = 200$	32.5%	45%	70%	75%	69.17%	74.17%	82.5%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	34.17%	43.33%	61.67%	60.83%	78.33%	80%	85.83%
$NoHN = 300$	30%	36.67%	55%	70.83%	75%	83.33%	81.67%
$NoHN = 400$	21.67%	30%	45%	55.83%	67.5%	64.17%	74.17%



## 6.8 Appendix 6.1

Table 6.12: Simulation 10: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Crowd Talking NSN with G.712 Type Handset at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the TIMIT Database

Simulation 10: The SIA for Crowd Talking NSN-WH for the <b>TIMIT Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	10%	25%	40.83%	50%	60%	73.33%	<b>85%</b>
CMVNMFCC( $\mathbf{f}_2$ )	4.17%	15%	30.83%	50%	63.33%	73.33%	77.5%
FWPNCC ( $\mathbf{g}_1$ )	4.17%	9.17%	20%	37.5%	45%	55%	55.83%
CMVNPNC ( $\mathbf{g}_2$ )	1.67%	10%	17.5%	30.83%	37.5%	50.83%	56.67%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	7.5%	24.17%	40%	50%	61.67%	75.83%	82.5%
$i_{WSF}$ at $\omega_2=0.8$	10%	19.17%	33.33%	55.83%	60%	76.67%	77.5%
$i_{WSF}$ at $\omega_3=0.77$	6.67%	18.33%	33.33%	55.83%	60%	72.5%	77.5%
$i_{WSF}$ at $\omega_4=0.7$	6.67%	18.33%	30%	54.17%	60%	72.5%	77.5%
$i_{Maximum}$	0.83%	10.83%	8.33%	27.5%	38.33%	39.17%	50.83%
$i_{Mean}$	4.17%	11.67%	12.5%	40%	49.17%	60.83%	67.5%
$i_{Cumulative}$	6.67%	10.83%	12.5%	44.17%	47.5%	58.33%	67.5%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	5.83%	15.83%	40%	57.5%	58.33%	76.67%	79.17%
$i_{interleaving}$ (2d) $NoHN = 200$	5%	15.83%	35.83%	56.67%	63.33%	75%	80%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	2.5%	15%	24.17%	54.17%	57.5%	62.5%	78.33%
$NoHN = 300$	2.5%	11.67%	37.5%	46.67%	65%	73.33%	76.67%
$NoHN = 400$	6.67%	10.83%	19.17%	32.5%	53.33%	58.33%	72.5%

## 6.8 Appendix 6.1

Table 6.13: Simulation 1: The Speaker Identification Accuracy (SIA) as a Function of the UBM Mixture Sizes {8, 16, 32, 64, 128, 256, 512 } for the I-vector Approach for 100, 200 and 400 Dimensions for Original Speech Recordings (OSR) without Handset at UBM Mixture Size 256 for the SITW Database

Simulation 11: The SIA for OSR Without Handset for the SITW Database							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
Feature based I-vector							
Without Fusion							
With 100 dimension							
$NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	47.5%	61.67%	77.5%	79.17%	85%	84.17%	83.33%
CMVNMFCC( $\mathbf{f}_2$ )	49.17%	68.33%	78.33%	82.5%	83.33%	82.5%	85%
FWPNCC ( $\mathbf{g}_1$ )	41.67%	63.33%	73.33%	81.67%	84.17%	85%	82.5%
CMVNPNC ( $\mathbf{g}_2$ )	1.67%	10%	17.5%	30.83%	37.5%	50.83%	56.67%
Fusion Decision	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_2)$
Fusion based I-vector							
With 100 dimension							
$NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	47.5%	65%	79.17%	82.5%	84.17%	81.67%	84.17%
$i_{WSF}$ at $\omega_2=0.8$	50.83%	68.33%	78.33%	84.17%	83.33%	81.67%	83.33%
$i_{WSF}$ at $\omega_3=0.77$	52.5%	69.17%	77.5%	84.17%	82.5%	80%	83.33%
$i_{WSF}$ at $\omega_4=0.7$	49.17%	63.33%	79.17%	80.83%	84.17%	80.83%	82.5%
$i_{Maximum}$	25.83%	44.17%	59.17%	71.67%	70%	73.33%	80%
$i_{Mean}$	46.67%	61.67%	74.17%	79.17%	81.67%	81.67%	81.67%
$i_{Cumulative}$	36.67%	49.17%	75.83%	78.33%	83.33%	80%	82.5%
Fusion based I-vector							
With 200 dimension							
$i_{Concatenated} (2d)$							
$NoHN = 200$	53.33%	72.5%	81.67%	81.67%	85.83%	83.33%	85%
$i_{interleaving} (2d)$							
$NoHN = 200$	10%	9.17%	12.5%	19.17%	31.67%	26.67%	27.5%
Fusion based I-vector							
With 400 dimension							
$i_{Concatenated} (4d)$							
$NoHN = 200$	52.5%	69.17%	80.83%	85%	83.33%	85.83%	82.5%
$NoHN = 300$	48.33%	65%	79.17%	80.83%	81.67%	85.83%	85.83%
$NoHN = 400$	39.17%	52.5%	71.67%	78.33%	77.5%	80.83%	80%

## 6.8 Appendix 6.1

Table 6.14: Simulation 2: The SIA for Different GMCs to the I-vector Approach for 100, 200 and 400 Dimensions Under AWGN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the SITW Database

Simulation 12: The SIA for AWGN-WH at 16 kHz for the <a href="#">SITW Database</a>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	5.83%	10.83%	20.83%	42.5%	66.67%	74.17%	80%
CMVNMFCC( $\mathbf{f}_2$ )	5.83%	15.83%	21.67%	42.5%	65%	74.17%	80%
FWPNCC ( $\mathbf{g}_1$ )	4.17%	11.67%	31.67%	58.33%	76.67%	78.33%	<b>84.17%</b>
CMVNPNC ( $\mathbf{g}_2$ )	4.17%	13.33%	35%	65.83%	79.17%	82.5%	82.5%
Fusion Decision	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	5%	14.17%	21.67%	45.83%	68.33%	75.83%	80.83%
$i_{WSF}$ at $\omega_2=0.8$	5.83%	12.5%	20%	46.67%	69.17%	75%	80%
$i_{WSF}$ at $\omega_3=0.77$	5.83%	12.5%	20%	48.33%	68.33%	77.5%	78.33%
$i_{WSF}$ at $\omega_4=0.7$	4.17%	12.5%	20.83%	45%	68.33%	75.83%	80%
$i_{Maximum}$	3.33%	7.5%	16.67%	45%	53.33%	68.33%	70%
$i_{Mean}$	5%	8.33%	25.83%	54.17%	67.5%	77.5%	78.33%
$i_{Cumulative}$	4.17%	7.5%	25%	54.17%	68.33%	76.67%	76.67%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	6.67%	16.67%	30.83%	56.67%	74.17%	79.17%	81.67%
$i_{interleaving}$ (2d) $NoHN = 200$	6.67%	5.83%	8.33%	17.5%	25%	26.67%	24.17%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	2.5%	13.33%	25.83%	59.17%	71.67%	77.5%	80%
$NoHN = 300$	2.5%	14.17%	32.5%	57.5%	70%	75%	82.5%
$NoHN = 400$	0.83%	7.5%	22.5%	46.67%	65.83%	73.33%	78.33%

## 6.8 Appendix 6.1

Table 6.15: Simulation 3: The SIA for Different GMCs to the I-vector Approach for 100, 200 and 400 Dimensions Under Street Traffic NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the SITW Database

Simulation 13: The SIA for Street Traffic NSN-WH at 16 kHz for the <a href="#">SITW Database</a>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	21.67%	33.33%	53.33%	65.83%	74.17%	81.67%	82.5%
CMVNMFCC( $\mathbf{f}_2$ )	19.17%	34.17%	48.33%	61.67%	74.17%	79.17%	81.67%
FWPNCC ( $\mathbf{g}_1$ )	5%	10%	23.33%	50.83%	70.83%	77.5%	83.33%
CMVNPNC ( $\mathbf{g}_2$ )	6.67%	14.17%	33.33%	54.17%	74.17%	81.67%	<b>84.17%</b>
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	22.5%	35.83%	54.17%	65%	75.83%	76.67%	80%
$i_{WSF}$ at $\omega_2=0.8$	20.83%	34.17%	51.67%	66.67%	75%	77.5%	80%
$i_{WSF}$ at $\omega_3=0.77$	20%	34.17%	51.67%	62.5%	73.33%	78.33%	79.17%
$i_{WSF}$ at $\omega_4=0.7$	20%	33.33%	52.5%	63.33%	72.5%	78.33%	81.67%
$i_{Maximum}$	10%	19.17%	29.17%	46.67%	61.67%	65.83%	73.33%
$i_{Mean}$	13.33%	23.33%	42.5%	55.83%	70.83%	76.67%	80.83%
$i_{Cumulative}$	13.33%	20.83%	43.33%	50.83%	71.67%	78.33%	81.67%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	12.5%	32.5%	53.33%	70.83%	78.33%	<b>84.17%</b>	82.5%
$i_{interleaving}$ (2d) $NoHN = 200$	6.67%	7.5%	13.33%	24.17%	23.33%	30%	19.17%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	2.5%	29.17%	47.5%	63.33%	74.17%	81.67%	79.17%
$NoHN = 300$	14.17%	35.83%	50.83%	65%	75%	78.33%	81.67%
$NoHN = 400$	12.5%	27.5%	50%	63.33%	75%	75.83%	75.83%

## 6.8 Appendix 6.1

Table 6.16: Simulation 4: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Bus Interior NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the SITW Database

Simulation 14: The SIA for Bus Interior-WH at 16 kHz for the <a href="#">SITW Database</a>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	65%	71.67%	75.83%	77.5%	80%	83.33%	80.83%
CMVNMFCC( $\mathbf{f}_2$ )	63.33%	70%	75.83%	78.33%	78.33%	80%	85%
FWPNCC ( $\mathbf{g}_1$ )	25.83%	46.67%	65%	75%	79.17%	80.83%	80.83%
CMVNPNC ( $\mathbf{g}_2$ )	30%	46.67%	65.83%	75%	80%	83.33%	84.17%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	63.33%	71.67%	75%	79.17%	79.17%	82.5%	84.17%
$i_{WSF}$ at $\omega_2=0.8$	63.33%	75.83%	73.33%	76.67%	78.33%	81.67%	82.5%
$i_{WSF}$ at $\omega_3=0.77$	64.17%	75%	72.5%	76.67%	78.33%	82.5%	81.67%
$i_{WSF}$ at $\omega_4=0.7$	60.83%	72.5%	72.5%	75.83%	77.5%	81.67%	80.83%
$i_{Maximum}$	36.67%	50.83%	55%	78.33%	70.83%	74.16%	80%
$i_{Mean}$	40%	60%	65.83%	76.67%	75.83%	81.67%	83.33%
$i_{Cumulative}$	43.33%	58.33%	66.67%	77.5%	75%	80%	82.5%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	64.17%	70%	77.5%	80%	81.67%	84.17%	<b>86.67%</b>
$i_{interleaving}$ (2d) $NoHN = 200$	23.33%	15.83%	25.83%	25%	27.5%	30%	28.33%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	57.5%	69.17%	75.85%	80.83%	82.5%	83.33%	83.33%
$NoHN = 300$	61.67%	67.5%	74.17%	80%	83.33%	83.33%	83.33%
$NoHN = 400$	50.83%	60.83%	70.83%	75.83%	76.67%	79.17%	80.83%

## 6.8 Appendix 6.1

Table 6.17: Simulation 5: The SIA for Different GMCs to the I-vector Approach for 100, 200 and 400 Dimensions Under Crowd Talking NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the SITW Database

Simulation 15: The SIA for Crowd Talking-WH at 16 kHz for the <a href="#">SITW Database</a>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	29.17%	45.83%	58.33%	71.67%	78.33%	80.83%	81.67%
CMVNMFCC( $\mathbf{f}_2$ )	26.67%	44.17%	59.17%	70.83%	76.67%	79.17%	80%
FWPNCC ( $\mathbf{g}_1$ )	5%	24.17%	45%	61.67%	76.67%	80.83%	83.33%
CMVNPNC ( $\mathbf{g}_2$ )	5.83%	25%	45%	61.67%	80%	84.16%	83.33%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	26.67%	45%	56.67%	72.5%	73.33%	81.67%	81.67%
$i_{WSF}$ at $\omega_2=0.8$	29.17%	46.67%	57.5%	70%	73.33%	80%	80.83%
$i_{WSF}$ at $\omega_3=0.77$	27.5%	46.67%	57.5%	70%	71.67%	78.33%	80%
$i_{WSF}$ at $\omega_4=0.7$	26.67%	44.17%	52.5%	66.67%	74.17%	79.17%	80.83%
$i_{Maximum}$	14.17%	27.5%	37.5%	54.17%	66.67%	71.67%	74.5%
$i_{Mean}$	19.17%	34.17%	49.17%	62.5%	71.67%	80.83%	82.5%
$i_{Cumulative}$	17.5%	32.5%	47.5%	65.83%	71.67%	81.67%	81.67%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	22.5%	46.67%	65.83%	78.33%	80.83%	85%	85%
$i_{interleaving}$ (2d) $NoHN = 200$	7.5%	9.17%	15%	25%	17.5%	17.5%	34.17%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	18.33%	36.67%	62.5%	73.33%	84.17%	85%	85%
$NoHN = 300$	22.5%	46.67%	61.67%	73.33%	81.67%	80.83%	81.67%
$NoHN = 400$	17.5%	37.5%	55.83%	65.83%	75%	78.33%	80%

## 6.8 Appendix 6.1

Table 6.18: Simulation 1: The Speaker Identification Accuracy as a Function of the UBM Mixture Sizes {8, 16, 32, 64, 128, 256, 512 } for the I-vector Approach for 100, 200 and 400 Dimensions for Original Speech Recordings (OSR) Without Handset at UBM Mixture Size 256 for the NIST 2008 Database

Simulation 16: The SIA for OSR Without Handset for the <b>NIST 2008 Database</b>							
Methods	Mix8	Mix16	Mix32	Mix64	Mix128	Mix256	Mix512
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	50%	54.17%	85.83%	89.17%	90.83%	94.17%	91.67%
CMVNMFCC( $\mathbf{f}_2$ )	51.67%	58.33%	80%	86.67%	95%	95%	91.67%
FWPNCC ( $\mathbf{g}_1$ )	34.17%	50%	75%	85%	85.83%	87.5%	85.83%
CMVNPNC ( $\mathbf{g}_2$ )	34.17%	50.83%	78.33%	84.17%	89.17%	93.33%	89.17%
Fusion Decision	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	60%	59.17%	75.83%	88.33%	91.67%	93.33%	95%
$i_{WSF}$ at $\omega_2=0.8$	50.83%	59.17%	75.83%	92.5%	89.17%	92.5%	94.17%
$i_{WSF}$ at $\omega_3=0.77$	44.17%	60.83%	70.83%	87.5%	90%	93.33%	92.5%
$i_{WSF}$ at $\omega_4=0.7$	49.17%	55%	77.5%	84.17%	87.5%	92.5%	91.67%
$i_{Maximum}$	11.67%	53.33%	41.67%	51.67%	70.83%	73.33%	80%
$i_{Mean}$	35%	58.33%	67.5%	73.33%	84.17%	89.17%	89.17%
$i_{Cumulative}$	23.33%	45%	59.17%	70%	84.17%	89.17%	88.33%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	50.83%	65%	81.67%	89.17%	95%	95.83%	94.17%
$i_{interleaving}$ (2d) $NoHN = 200$	47.5%	55%	81.67%	90.83%	93.33%	<b>96.67%</b>	95%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	46.67%	53.33%	83.33%	86.66%	91.67%	92.5%	95%
$NoHN = 300$	39.17%	52.5%	75.83%	86.67%	86.67%	90.83%	89.17%
$NoHN = 400$	25%	42.5%	65%	80.83%	87.5%	88.33%	87.5%

## 6.8 Appendix 6.1

Table 6.19: Simulation 2: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under AWGN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the NIST 2008 Database

Simulation 17: The SIA Under AWGN-WH for NIST 2008 for the <b>NIST 2008 Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	1.67%	3.33%	9.17%	18.33%	34.17%	43.33%	48.33%
CMVNMFCC( $\mathbf{f}_2$ )	1.67%	3.33%	6.67%	20.83%	36.67%	44.17%	52.5%
FWPNCC ( $\mathbf{g}_1$ )	0.83%	0.83%	1.67%	7.5%	16.67%	33.33%	52.5
CMVNPNC ( $\mathbf{g}_2$ )	1.67%	1.67%	3.33%	5%	17.5%	33.33%	48.33%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_1)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	1.67%	3.33%	9.16%	21.67%	38.33%	45.84%	48.33%
$i_{WSF}$ at $\omega_2=0.8$	1.67%	3.33%	8.33%	16.67%	30.83%	40%	51.67%
$i_{WSF}$ at $\omega_3=0.77$	1.67%	3.33%	8.33%	18.33%	32.5%	40.83%	53.33%
$i_{WSF}$ at $\omega_4=0.7$	1.67%	3.33%	11.67%	15.83%	35.83%	40.83%	55%
$i_{Maximum}$	0.83%	0.83%	5.83%	7.5%	19.17%	18.33%	36.67%
$i_{Mean}$	1.67%	2.5%	11.67%	13.33%	15%	25%	45%
$i_{Cumulative}$	0.83%	2.5%	13.33%	12.5%	20.83%	23.33%	45%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	0.83%	1.67%	1.67%	56.67%	74.17%	79.17%	<b>81.67%</b>
$i_{interleaving}$ (2d) $NoHN = 200$	0.83%	1.67%	6.67%	10.83%	12.5%	7.5%	9.17%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	0.83%	2.5%	5.83%	7.5%	19.17%	40%	54.17%
$NoHN = 300$	1.67%	0.83%	1.67%	10.83%	21.67%	40%	52.5%
$NoHN = 400$	0.83%	0.83%	1.67%	4.17%	10.83%	32.5%	37.5%



## 6.8 Appendix 6.1

Table 6.20: Simulation 3: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Street Traffic NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the NIST 2008 Database

Simulation 18: The SIA for Street Traffic NSN-WH for the <b>NIST 2008 Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	4.17%	10%	21.67%	31.67%	37.5%	47.5%	51.67%
CMVNMFCC( $\mathbf{f}_2$ )	2.5%	12.5%	16.67%	26.67%	41.67%	54.17%	57.5%
FWPNCC ( $\mathbf{g}_1$ )	1.67%	11.67%	17.5%	37.5%	63.33%	73.33%	<b>78.33</b>
CMVNPNC ( $\mathbf{g}_2$ )	1.67%	3.33%	17.5%	43.33%	53.33%	71.67%	75.83%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	4.17%	14.17%	21.67%	29.17%	40.83%	51.67%	57.5%
$i_{WSF}$ at $\omega_2=0.8$	5%	13.33%	19.17%	30.83%	39.17%	58.33%	59.16%
$i_{WSF}$ at $\omega_3=0.77$	8.33%	12.5%	17.5%	27.5%	41.67%	60.83%	59.17%
$i_{WSF}$ at $\omega_4=0.7$	5%	13.33%	16.67%	26.67%	39.17%	57.5%	55.83%
$i_{Maximum}$	2.5%	4.17%	10%	21.67%	27.5%	64.6%	54.17%
$i_{Mean}$	5.83%	9.17%	19.17%	29.17%	45%	67.5%	70%
$i_{Cumulative}$	5%	8.33%	17.5%	28.33%	43.33%	68.33%	72.5%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	0.83%	5%	15.83%	28.33%	39.17%	52.5%	60.83%
$i_{interleaving}$ (2d) $NoHN = 200$	4.17%	6.67%	10%	10%	9.17%	14.17%	23.33%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	2.5%	13.33%	13.33%	39.17%	45.83%	70.83%	64.17%
$NoHN = 300$	3.33%	5.83%	17.5%	32.5%	52.5%	59.17%	63.33%
$NoHN = 400$	1.67%	1.67%	6.67%	18.33%	32.5%	66.67%	57.5%

## 6.8 Appendix 6.1

Table 6.21: Simulation 4: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Bus-Interior NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the NIST 2008 Database

Simulation 19: The SIA for Bus-Interior NSN-WH for the NIST 2008 Database							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	31.67%	32.5%	40.83%	44.17%	61.67%	71.67%	85%
CMVNMFCC( $\mathbf{f}_2$ )	25.83%	30.83%	42.5%	47.5%	56.67%	75.83%	86.67%
FWPNCC ( $\mathbf{g}_1$ )	10%	19.17%	37.5%	59.17%	77.5%	80%	85%
CMVNPNC ( $\mathbf{g}_2$ )	9.17%	20%	38.33%	60.38%	77.5%	84.17%	82.5%
Fusion Decision	$(\mathbf{f}_1, \mathbf{g}_1)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_1, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_1)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	26.67%	36.67%	41.67%	48.33%	55.83%	74.17%	83.33%
$i_{WSF}$ at $\omega_2=0.8$	21.67%	34.17%	44.17%	50%	58.33%	75%	87.5%
$i_{WSF}$ at $\omega_3=0.77$	19.17%	32.5%	45%	50.83%	57.5%	71.67%	87.5%
$i_{WSF}$ at $\omega_4=0.7$	17.5%	33.33%	41.67%	50%	55%	68.33%	85%
$i_{Maximum}$	11.67%	14.17%	30%	34.17%	53.33%	58.33%	69.17%
$i_{Mean}$	15%	30.83%	37.5%	47.5%	56.67%	73.33%	82.5%
$i_{Cumulative}$	13.33%	25%	35%	43.33%	61.67%	73.33%	84.17%
Fusion based I-vector With 200 dimension							
$i_{Concatenated} (2d)$ $NoHN = 200$	20.83%	25%	33.33%	54.17%	62.5%	77.5%	87.5%
$i_{interleaving} (2d)$ $NoHN = 200$	6.67%	8.33%	15%	17.5%	15%	17.5%	25%
Fusion based I-vector With 400 dimension							
$i_{Concatenated} (4d)$ $NoHN = 200$	20.83%	36.67%	41.67%	52.5%	80.83%	74.17%	85%
$NoHN = 300$	23.33%	23.33%	40.83%	55.83%	70%	79.17%	83.33%
$NoHN = 400$	16.67%	22.5%	32.5%	43.33%	53.33%	67.5%	79.17%

## 6.8 Appendix 6.1

Table 6.22: Simulation 5: The SIA for Different GMCs for the I-vector Approach for 100, 200 and 400 Dimensions Under Crowd Talking NSN with G.712 Type Handset (WH) at 16 kHz at Different SNR Levels  $\{0, 5, 10, 15, 20, 25, 30\}$  dB at UBM Mixture Size 256 for the NIST 2008 Database

Simulation 20: The SIA for Crowd Talking NSN-WH for the <b>NIST 2008 Database</b>							
Methods	0dB	5dB	10dB	15dB	20dB	25dB	30dB
Feature based I-vector Without Fusion							
With 100 dimension $NoHN = 100$							
FWMFCC ( $\mathbf{f}_1$ )	1.67%	4.17%	6.67%	20%	43.33%	55.83%	66.67%
CMVNMFCC( $\mathbf{f}_2$ )	4.17%	14.17%	23.33%	35%	45.83%	58.33%	70.83%
FWPNCC ( $\mathbf{g}_1$ )	5%	20.83%	25.83%	50%	63.33%	78.33%	82.5
CMVNPNC ( $\mathbf{g}_2$ )	2.5%	10%	26.67%	40%	68.33%	80.83%	<b>85%</b>
Fusion Decision	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_1)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$	$(\mathbf{f}_2, \mathbf{g}_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF}$ at $\omega_1=0.9$	5.83%	13.33%	21.67%	32.5%	50%	58.33%	75%
$i_{WSF}$ at $\omega_2=0.8$	5.83%	14.17%	20.83%	32.5%	48.33%	58.33%	75%
$i_{WSF}$ at $\omega_3=0.77$	5.83%	16.67%	20%	30%	47.5%	57.5%	75.83%
$i_{WSF}$ at $\omega_4=0.7$	4.17%	13.33%	25%	30.83%	50%	57.5%	74.17%
$i_{Maximum}$	5.83%	6.67%	12.5%	30%	31.67%	48.33%	60.83%
$i_{Mean}$	5%	12.5%	17.5%	36.67%	48.33%	62.5%	74.17%
$i_{Cumulative}$	5%	9.17%	19.17%	38.33%	51.67%	63.33%	78.33%
Fusion based I-vector With 200 dimension							
$i_{Concatenated}$ (2d) $NoHN = 200$	5.83%	11.67%	23.33%	30.83%	47.5%	60.83%	67.5%
$i_{interleaving}$ (2d) $NoHN = 200$	6.67%	2.5%	14.17%	10.83%	11.67%	18.33%	15.83%
Fusion based I-vector With 400 dimension							
$i_{Concatenated}$ (4d) $NoHN = 200$	4.17%	19.16%	21.67%	26.67%	50%	70.83%	79.17%
$NoHN = 300$	2.5%	20.83%	15.83%	30%	54.17%	70%	77.5%
$NoHN = 400$	0.83%	3.33%	4.16%	10.83%	42.5%	55%	55.83%

## 6.8 Appendix 6.1

Table 6.23: Simulation 1: The Speaker Identification Accuracy as a Function of the UBM Mixture Sizes {8, 16, 32, 64, 128, 256, 512 } for the NTIMIT Database

Simulation 21: Results For <b>the NTIMIT Database</b>							
<b>I-vector Methods</b>	<b>Mix=8</b>	<b>Mix=16</b>	<b>32</b>	<b>64</b>	<b>128</b>	<b>256</b>	<b>Mix=512</b>
Feature based I-vector Without Fusion With 100 dimension $NoHN = 100$							
$i_{FW}$	6.67%	5%	15.83%	22.5%	36.67%	37.5%	40%
$i_{CMVN}$	5.33%	5%	14.67%	20.33%	34.33%	35.83%	38.67%
$\ddot{i}_{FW}$	1.67%	10%	25.83%	29.17%	29.17%	35%	34.17%
$\ddot{i}_{CMVN}$	3.33%	8.33%	20%	35%	36.67%	39.17%	40%
Fusion Decision	$(f_1, g_2)$	$(f_1, g_1)$	$(f_1, g_1)$	$(f_1, g_2)$	$(f_1, g_2)$	$(f_1, g_2)$	$(f_1, g_2)$
Fusion based I-vector With 100 dimension $NoHN = 100$							
$i_{WSF} \omega_1=0.9$	0.83%	7.5%	20%	25.83%	34.17%	37.5%	<b>42.5%</b>
$i_{WSF} \omega_2=0.8$	3.33%	5.83%	20.83%	21.67%	31.67%	37.5%	39.17%
$i_{WSF} \omega_3=0.77$	2.5%	3.33%	15.83%	26.67%	32.5%	36.67%	38.33%
$i_{WSF} \omega_4=0.7$	3.33%	6.67%	15%	20%	30%	35.83%	35.83%
$i_{Maximum}$	1.67%	2.5%	6.67%	9.17%	14.17%	20%	19.17%
$i_{Mean}$	3.33%	5%	17.5%	20%	23.33%	27.5%	31.67%

# Chapter 7

## Conclusions and Future Work

In this chapter, there are three sections. Section 7.1 summarizes the contributions of this thesis, Section 7.2 covers the conclusion, and Section 7.3 has suggestions for future work.

### 7.1 Contributions Overview

This section presents the main contributions of this thesis as covered in chapters four, five and six, respectively.

In Chapter 4, a closed set text independent speaker identification system was established. Four main simulations using a fixed original speech recordings length (129,250 samples 8 seconds in length) were performed to calculate the speaker identification accuracy for different Gaussian mixture components and different feature dimensions based on different fusion techniques. These fusion methods are: late fusion (score-based 16 Feature Dimension FD), early fusion (feature-based 32 FD) and early-late fusion (feature score-based 32 FD), concatenated static and dynamic features (feature-based 39FD), and finally the multiplication of scores independent of different feature dimensions (16, 32 and 39).

In Chapter 5, a comprehensive evaluation based on GMM-UBM approach was provided of text independent closed set speaker identification in the presence of AWGN and NSN types with a G.712 type handset at 16 kHz, to provide benchmark evaluations of three different databases for other researchers working in this area. It was found that the NIST 2008 database seems to have the best performance in the evaluation without noise and handset. However, it appears to be very sensitive and achieved the worst performance compared with other

databases used, in the presence of AWGN with handset and for NSN types with handset, at ranges of SNRs (0-15) dB. In addition, the TIMIT database showed the second best performance in both original speech recordings speech and AWGN. However, the new SITW database had less reduction in SIA compared with the TIMIT and the NIST 2008 databases in terms of AWGN and NSN over the range (0-15) dB. Fusion scores using equal weighting between the MFCC and the PNCC features represented by the fusion mean method was the best fusion method in both original speech recordings and NSN, and sometimes for AWGN, while for a constant spectrum of noise as in AWGN it seems the maximum fusion approach mitigated the reduction in SIA much better.

In Chapter 6, four feature combinations with seven fusion methods based on I-vector were investigated to develop a novel closed set text independent speaker identification system. The new system was modelled with fusion-based multi-dimensional I-vectors and classified with a single layer neural network using ELM. This has not been used before for speaker identification purposes. The system was tested using four different databases: TIMIT, NTIMIT, SITW and NIST 2008 databases, with 120 speakers from each database (total 480 speakers, 4,800 speech utterances). The fusion techniques were used to improve the SIA in original speech recordings and improve the reduction in SIA in the presence of noise and handset. This chapter can be summarized by the following points. Firstly, the identification accuracy for the I-vector seems to outperform the GMM-UBM for most environments with the SITW and NIST 2008 databases. However, in the TIMIT database, the system outperformed GMM-UBM techniques for original speech recordings, and also outperformed under AWGN without handset. It also seemed better for some SNR levels with street and crowd talking. In contrast, for bus interior NSN, the GMM-UBM achieved less reduction in SIA compared with the I-vector approach. Additionally, fusion techniques may mitigate the reduction caused by different noise environments and the handset effect, whereas fusion weights generally seem to be the best of all feature and fusion methods used. However, the new database using SITW demonstrated that the identification accuracy achieved by the I-vector approach was better than the corresponding results for GMM-UBM method for all challenging environments. With the NIST 2008 database, it seems the output from GMM-UBM is better than small mixture sizes with the I-vector, while this result was reversed when the

UBM mixture size was increased, to give slightly better results in original speech recordings. The I-vector generally gives higher SIA than GMM-UBM for different noise types, except for the bus NSN, where the GMM-UBM outperforms the I-vector. With the TIMIT database, the I-vector approach has better performance than GMM-UBM in original speech recordings and for AWGN without handset, while in other types of NSN the I-vector outperforms GMM-UBM on some SNR levels.

Secondly, this chapter also suggests a simple, efficient ELM classifier, which no one has yet used for this combination of features with the I-vector for speaker identification. Thirdly, the smallest I-vector with 100 and 200 dimensions has higher SIA compared with other I-vector dimensions with 400 and 800. Fourthly, almost the best SIAs are achieved at UBM mixture size 256 for original speech recordings, while in noise and handset conditions the best SIA achieved at SNR 30 dB for all databases used. Fifthly, in noisy conditions, the worst SIA was achieved at AWGN due to the stationary spectrum for the noise, while the highest SIA was obtained for bus interior NSN followed by the street NSN, then by crowd talking NSN. Finally, the best fusion type for the best SIA is considered in this chapter. Then there follows the best fusion method for the various environments, such as original speech recordings, AWGN-WO/WH, street NSN-WO/WH, bus NSN-WO/WH, crowd talking NSN-WO/WH. With the TIMIT database, the best fusion methods for the highest SIA is the weighted sum fusion, while in the SITW and NIST 2008 databases the best fusion method, according to the best SIA, is the fusion for concatenated-2d. In addition, some other fusion types are also useful to achieve improvements in SIA for different environments and different SNR levels.

## 7.2 Conclusions

This thesis includes comprehensive experiments, simulations and evaluations, and, therefore, this section will focus on the most significant findings concerning the best methods and approaches, as well as the improvements achieved in this thesis. Firstly, this section presents three main tables which show the improvements found in using the I-vector approach, compared with the GMM-UBM for the three databases, TIMIT, SITW and NIST 2008 databases, as explained in the tables from Table 7.1 to Table 7.3. In addition, 120 speakers from each database were

## 7.2 Conclusions

---

selected using a total of 360 speakers and 3,600 speech utterances. In this thesis, four databases were used with 480 speakers in total (4,800 speech utterances), and a small experiment was also completed using the NTIMIT database. This section emphasises only the three databases. Table 7.1 shows the improvements of the I-vector approach compared with the GMM-UBM for the TIMIT database under original speech recordings, AWGN With Handset (WH), street traffic NSN-WH, bus interior NSN-WH and crowd talking NSN-WH, represented by parts (a), (b), (c), (d) and (e) from Table 7.1, respectively. The best SIA is selected per mixture size or per SNR level for the I-vector, regardless of feature combination type, fusion type and I-vector dimensions (100, 200 and 400), and then the corresponding SIA for the GMM-UBM are selected. It can be observed that there was no improvements in parts (d) and (e) and the GMM-UBM; however, the major improvements are achieved for part (b) at AWGN-WH as explained in green highlights. Also, for original speech recordings part (a), the I-vector attained the highest SIA with 96.67 % at mixture size 256 with 2.65 % improvement over the GMM-UBM. Table 7.2 is the most important table, which shows that the new database (SITW 2016) has significant improvements for the I-vector approach over the GMM-UBM in all challenging environments. However, very few SNR levels in parts (a), (d) and (e) for the GMM-UBM have better accuracy than the I-vector. Table 7.3 illustrates the NIST 2008 improvements, and is considered the second database after the SITW in terms of the I-vector improvements. Only a few SNR levels for noisy speech in parts (c), (d) and (e) have an accuracy for the GMM-UBM that outperforms the corresponding point in the I-vector method. Moreover, the I-vector with small mixture sizes has lower SIA for the I-vector compared with the GMM-UBM for original speech recordings evaluation. Secondly, the best feature combination for the I-vector with and without fusion based I-vector according to each database used, can be seen in Table 7.4 and Table 7.5 which are proposed for this section. where the symbols **F1**, **F2**, **F3**, **F4** represent the feature based methods (without fusion) namely FWMFCC, CMVNMFCC, FWPNCC and CMVNPNC, respectively. Furthermore, the symbols **F5**, **F6**, and **F7** represent the feature based methods with weighted sum, maximum and mean fusion, respectively. In addition, **F8**, **F9**, **F10**, and **F11** represent cumulative, concatenated-2d, interleaving-2d and concatenated-4d. In Table 7.4, the best feature with and without fusion based I-vector is selected



## 7.2 Conclusions

---

according to the highest SIA for each environment with the three databases (TIMIT, SITW and NIST 2008). According to Table 7.4, it is clear that the fusion-based I-vector has better SIA compared with feature combination based without the fusion; thereby, the fusion has a significant effect on improving the SIA. Furthermore, the weighted sum fusion and the concatenated fusion have the best fusion methods which work to improve the SIA in the I-vector approach.

## 7.2 Conclusions

Table 7.1: Percentage Improvements for the I-vector Approach Compared with the GMM-UBM Approach for the TIMIT Database Under Different Environments

(a) TIMIT Database for Original Speech Recordings							
Approaches	Mix 8	Mix 16	Mix 32	Mix 64	Mix 128	Mix 256	Mix 512
GMM-UBM	80.83%	84.17%	90%	93.33%	94.17%	94.17%	95%
I-vector	44.16%	65.83%	87.5%	90%	94.17%	96.67%	95%
Improvement	NA	NA	NA	NA	0%	2.65%	0%

(b) TIMIT Database for AWGN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	2.5%	4.17%	7.5%	20%	39.17%	51.67%	75.83%
I-vector	5%	16.67%	33.33%	51.67%	55.83%	68.33%	74.16%
Improvement	100%	299.76%	344.4%	158.35%	42.53%	32.24%	NA

(c) TIMIT Database for Street Traffic NSN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	6.67%	18.33%	31.67%	55%	74.17%	84.17%	90%
I-vector	16.67%	30%	38.33%	55%	66.67%	75.83%	82%
Improvement	149.93%	63.67%	21.03%	0%	NA	NA	NA

(d) TIMIT Database for Bus Interior NSN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	56.67%	72.5%	83.33%	85.83%	90.83%	91.67%	91.67%
I-vector	48.33%	64.17%	74.17%	79.17%	81.67%	85%	89.17%
Improvement	NA	NA	NA	NA	NA	NA	NA

(e) TIMIT Database for Crowd Talking NSN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	10%	19.17%	39.17%	62.5%	74.17%	84.17%	90%
I-vector	10%	25%	40.83%	57.5%	65%	76.67%	85%
Improvement	NA	NA	NA	NA	NA	NA	NA

	Highlighted the SIA to the I-vector and GMM-UBM When Both are Equal
	Highlighted the SIA to the I-vector and GMM-UBM When SIA of I-vector Outperform GMM-UBM
	Highlighted the Percentage Improvement of the SIA to the I-vector Compared with the GMM-UBM

## 7.2 Conclusions

Table 7.2: Percentage Improvements for the I-vector Approach Compared with the GMM-UBM Approach for the SITW Database Under Different Environments

(a) SITW Database for Original Speech Recordings							
Approaches	Mix 8	Mix 16	Mix 32	Mix 64	Mix 128	Mix 256	Mix 512
GMM-UBM	73.33%	76.67%	78.33%	79.17%	80.83%	80.83%	82.5%
I-vector	53.33%	72.5%	81.67%	84.17%	85.83%	85.83%	85.83%
Improvement	NA	NA	4.26%	6.32%	6.19%	6.19%	4.03%

(b) SITW Database for AWGN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	4.17%	10.83%	23.33%	53.33%	74.17%	78.33%	78.33%
I-vector	6.67%	16.67%	35%	65.83%	79.17%	82.5%	84.17%
Improvement	59.95%	53.92%	50.02%	23.44%	6.74%	5.32%	7.46%

(c) SITW Database for Street Traffic NSN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	15.83%	24.17%	46.88%	63.33%	74.17%	79.17%	81.67%
I-vector	22.5%	35.83%	54.17%	70.83%	78.33%	84.17%	84.17%
Improvement	42.13%	48.24%	15.55%	11.84%	5.61%	6.32%	3.06%

(d) SITW Database for Bus Interior NSN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	66.67%	72.5%	75%	79.17%	80%	81.67%	80.83%
I-vector	65%	75.83%	77.5%	80.83%	83.33%	84.17%	86.67%
Improvement	NA	4.59%	3.33%	2.1%	4.16%	3.06%	7.23%

(e) SITW Database for Crowd Talking NSN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	20%	65%	53.33%	72.5%	75%	78.33%	82.5%
I-vector	29.17%	46.67%	65.83%	78.33%	84.17%	85%	85%
Improvement	45.85%	NA	NA	8.04%	12.23%	8.52%	3.03%

- Highlighted the SIA to the I-vector and GMM-UBM When Both are Equal
- Highlighted the SIA to the I-vector and GMM-UBM When SIA of I-vector Outperform GMM-UBM
- Highlighted the Percentage Improvement of the SIA to the I-vector Compared with the GMM-UBM

## 7.2 Conclusions

Table 7.3: Percentage Improvements for the I-vector Approach Compared with the GMM-UBM Approach for the NIST 2008 Database Under Different Environments

(a) NIST 2008 Database for Original Speech Recordings							
Approaches	Mix 8	Mix 16	Mix 32	Mix 64	Mix 128	Mix 256	Mix 512
GMM-UBM	91.67%	91.67%	94.17%	95.83%	95%	95%	95%
I-vector	60%	65%	85.83%	92.5%	95%	96.67%	95%
Improvement	NA	NA	NA	NA	0%	1.76%	0%

(b) NIST 2008 Database for AWGN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	0.83%	2.5%	3.33%	9.16%	15.83%	21.67%	26.67%
I-vector	1.67%	3.33%	13.33%	56.67%	74.17%	79.17%	81.67%
Improvement	101.2%	33.2%	300.3%	518.67%	368.54%	265.34%	206.22%

(c) NIST 2008 Database for Street Traffic NSN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	1.67%	5.83%	15%	34.17%	55.83%	74.17%	80%
I-vector	8.33%	14.17%	21.67%	43.33%	63.33%	73.33%	78.33%
Improvement	398.8%	143.06%	44.46%	26.8%	13.43%	NA	NA

(d) NIST 2008 Database for Bus Interior NSN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	22.5%	32.5%	45.83%	58.33%	75%	86.67%	92.5%
I-vector	31.67%	36.67%	45%	60.38%	80.83%	84.17%	87.5%
Improvement	40.76%	62.98%	NA	3.51%	7.77%	NA	NA

























(e) NIST 2008 Database for Crowd Talking NSN with G.712 Type Handset at 16 kHz							
Approaches	SNR 0dB	SNR 5dB	SNR 10dB	SNR 15dB	SNR 20dB	SNR 25dB	SNR 30dB
GMM-UBM	10%	15.83%	30%	45.83%	68.33%	79.17%	84.17%
I-vector	6.67%	20.83%	26.67%	50%	68.33%	80.83%	85%
Improvement	NA	31.59%	NA	9.1%	0%	2.1%	0.99%

	Highlighted the SIA to the I-vector and GMM-UBM When Both are Equal
	Highlighted the SIA to the I-vector and GMM-UBM When SIA of I-vector Outperform GMM-UBM
	Highlighted the Percentage Improvement of the SIA to the I-vector Compared with the GMM-UBM

In Table 7.5, the best feature with and without fusion method based GMM-UBM approach are presented according to the highest SIA for each speech condition for the TIMIT, SITW and NIST 2008 databases. Likewise in Table 7.4, it is evident that the fusion-based GMM-UBM in Table 7.5 also outperforms the corresponding feature combination for the same approach. However, both fusion mean and the

## 7.2 Conclusions

Table 7.4: The Best Feature With and Without Fusion I-vector Methods According to the Highest SIA for Three Databases

Environments	Feature Based (Without Fusion)				Feature Based (With Fusion)						
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
Original databases recordings					 Table 6.3				 Table 6.13	 Table 6.18	 Table 6.13
AWGN-WOH at 30 dB									 Table 6.5	 Table 6.5	
AWGN-WH at 30 dB			 Table 6.14				 Table 6.9		 Table 6.19		
Street NSN-WOH at 30 dB					 Table 6.6						
Street NSN-WH at 30 dB			 Table 6.20	 Table 6.14	 Table 6.10						
Bus NSN-WOH at 30 dB										 Table 6.7	
Bus NSN-WH at 30 dB	 Table 6.11				 Table 6.11	 Table 6.21			 Table 6.16	 Table 6.21	
Crowd talking – NSN WOH at 30 dB									 Table 6.8		
Crowd-talking- NSN WH at 30 dB	 Table 6.12			 Table 6.22					 Table 6.17		 Table 6.17

Where: F1: FWMFCC F2: CMVNMFCC F3: FWPNCC F4: CMVNPNC F5: Weighted Sum  
F6: Maximum F7: Mean F8: Cumulative F9: Concatenated-2d F10: Interleaving-2d  
F11: Concatenated-4d.

WOH Without Handset

WH With G.712 Type Handset at 16 kHz



This symbol for TIMIT-Database



This symbol for SITW-Database






















This symbol for NIST 2008-Database

weighted sum have the best fusion methods. It is clear from this figure that the best method for all databases without any noise and handset conditions was the weighted sum fusion, which is represented by **F5**. However, in noisy conditions mean fusion which is represented by **F7** achieved the best SIAs for all databases.

## 7.2 Conclusions

Table 7.5: Best SIA Performance for Feature Based Speaker Identification With and Without Fusion for Three Databases Using GMM-UBM Approach

Environments	Feature Based (Without Fusion)				Feature Based (With Fusion)		
	F1	F2	F3	F4	F5	F6	F7
Original databases recordings without additional noise and handset effects Results were taken from Table 5.2					  		
AWGN-WH Results were taken from Table 5.3			 30 dB	 30 dB			   30 dB 25 dB 30 dB
Street NSN-WH Results were taken from Table 5.4			 30 dB				  30 dB 30 dB
Bus NSN-WH Results were taken from Table 5.5		 30 dB			  30 dB 25 dB		  25 dB 30 dB
Crowd-talking-NSN WH Results were taken from Table 5.6					 30 dB		  30 dB 30 dB

where: F1: FWMFCC F2: CMVNMFCC F3: FWPNCC F4: CMVNPNC F5: Weighted Sum  
F6: Maximum F7: Mean

WH With G.712 Type Handset at 16 kHz



This Symbol for TIMIT-Database



This Symbol for SITW-Database



This Symbol for NIST 2008-Database

### 7.3 Suggestions for future work

Suggestions for future work are summarized by the following bullet points:

- Student's-t Mixture Model instead of Gaussian mixture model is suggested as alternative for future work, due to that having heavier tails than the corresponding Gaussian distribution, which can improve the performance accuracy.
- According to the I-vector approach, the work for this thesis can be extended to other applications, such as emotion and language recognition based on the I-vector, and the NIST 2008 database already has different languages; only English speakers were exploited in this thesis.
- Speaker I-vector Machine Learning Challenge as well as the Language I-vector Machine Learning Challenge can be exploited to develop a new I-vector speaker and language identification system for future work, perhaps to compare the results with a new identification system.
- This thesis opens the door for other biometric recognitions to exploit total variability space, so instead of using the speaker and channel variabilities based on voice-print or speech biometrics, it might be used to improve two different variabilities based on image processing, and then implement their total variability space for Iris, fingerprint, finger texture, face, or any other biometrics applications.
- Increasing the number of speakers can be suggested, while different channel variabilities can be considered as well. In addition, a telephone channel from NIST 2008 can be exploited with the new simulations to implement or compare the results with the new system. In addition, a new database could be employed, such as: the GRID audiovisual, the MOBIO, VoxForge and MOCHA TIMIT databases.
- This thesis gives new suggestions for fusion systems, which might be useful for other biometrics recognition, for example using maximum, cumulative, interleaving and concatenated fusion methods for multi-features, as well as those classical fusion methods using mean and weighted sum.

### 7.3 Suggestions for future work

---

- The Extreme Learning Machines have been shown in this thesis to be simple, powerful and efficient compared with other traditional classification methods, and, therefore, they can be exploited for other works based on other biometrics applications. Furthermore, the current thesis encourages other researchers working in biometrics applications to exploit the characteristics for the ELM to improve their systems.
- A new speaker identification system can be developed for multi-speakers/multi-channels instead of the single speakers/single channel used in this thesis. This can be achieved by exploiting both NIST 2008 and the new database SITW 2016, where both databases contain multi-speakers.
- A very recent new database “ Noisy TIMIT Speech” appeared after the generation of the work in this thesis in 2017 and it would be interesting to use this for comparative evaluations.



# References

- [1] R. S. S. Kumari, S. S. Nidhyananthan, and A. G, “Fused MEL feature sets based text-independent speaker identification using Gaussian mixture model,” *Procedia Engineering*, vol. 30, pp. 319–326, 2012.
- [2] A. K. Jain, A. Ross, and S. Prabhakar, “An introduction to biometric recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.
- [3] R. Togneri and D. Pullella, “An overview of speaker identification: Accuracy and robustness issues,” *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011.
- [4] J. Benesty, *Springer Handbook of Speech Processing*. Springer Science & Business Media, 2008.
- [5] S. S. Vyawahare, “Speaker recognition: A review,” in *International Journal of Engineering Research and Technology*, vol. 2, no. 2 (February-2013). ESRSA Publications, 2013.
- [6] A. Rashed and W. M. Bahgat, “Modified technique for speaker recognition using ANN,” *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 13, no. 8, pp. 8–13, 2013.
- [7] S. J. Arora and R. P. Singh, “Automatic speech recognition: a review,” *International Journal of Computer Applications*, vol. 60, no. 9, pp. 34–44, 2012.
- [8] N. Alcaraz Meseguer, “Speech analysis for automatic speech recognition,” Master’s thesis, Department for Electronics and Telecommunications, Norwegian University of Science and Technology, 2009.

- [9] “The human ear.” [Online]. Available: <https://www.cmft.nhs.uk/media/327153/adultcochlearimpantprogrammebooklet.pdf>
- [10] “Cochlear implant.” [Online]. Available: <https://cochlearimplanthelp.com/journey/choosing-a-cochlear-implant/electrodes-and-channels/>
- [11] H. Rask-Andersen, W. Liu, E. Erixon, A. Kinnefors, K. Pfaller, A. Schrott-Fischer, and R. Glueckert, “Human cochlea: anatomical characteristics and their relevance for cochlear implantation,” *The Anatomical Record*, vol. 295, no. 11, pp. 1791–1811, 2012. [Online]. Available: <http://onlinelibrary.wiley.com/doi/10.1002/ar.22599/pdf>
- [12] A. Maesa, F. Garzia, M. Scarpiniti, and R. Cusani, “Text independent automatic speaker recognition system using MEL-frequency cepstrum coefficient and Gaussian mixture models,” *Journal of Information Security*, vol. 3, no. 04, pp. 335–340, 2012.
- [13] G. Nijhawan and M. Soni, “A new design approach for speaker recognition using MFCC and VAD,” *International Journal of Image, Graphics and Signal Processing*, vol. 5, no. 9, pp. 43–49, 2013.
- [14] C. Kim and R. M. Stern, “Power-normalized cepstral coefficients PNCC for robust speech recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4101–4104, 2012.
- [15] —, “Power-normalized cepstral coefficients PNCC for robust speech recognition,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 7, pp. 1315–1329, 2016.
- [16] E. Ambikairajah, J. M. K. Kua, V. Sethu, and H. Li, “PNCC-Ivector-SRC based speaker verification,” *Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1–7, 2012.
- [17] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” *The Speaker Recognition Workshop Crete, Greece*, 2001.
- [18] H. Beigi, *Fundamentals of Speaker Recognition*. Springer Science & Business Media, 2011.

- [19] N. V. Prasad and S. Umesh, “Improved cepstral mean and variance normalization using Bayesian framework,” *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 156–161, 2013.
- [20] S. O. Sadjadi, M. Slaney, and L. Heck, “MSR identity toolbox: A Matlab toolbox for speaker-recognition research,” *Speech and Language Processing Technical Committee Newsletter*, vol. 1, no. 4, 2013.
- [21] A. Reda and B. Aoued, “Artificial neural network & MEL-frequency cepstrum coefficients-based speaker recognition,” *3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications–Tunisia*, vol. 76, pp. 89–93, 2005.
- [22] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, “Speaker identification by combining MFCC and phase information in noisy environments,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4502–4505, 2010.
- [23] Y. Yujin, Z. Peihua, and Z. Qun, “Research of speaker recognition based on combination of LPCC and MFCC,” *IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*, vol. 3, pp. 765–767, 2010.
- [24] P. K. Ajmera, D. V. Jadhav, and R. S. Holambe, “Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram,” *Pattern Recognition*, vol. 44, no. 10, pp. 2749–2759, 2011.
- [25] E. B. Tazi, A. Benabbou, and M. Harti, “Efficient text independent speaker identification based on GFCC and CMN methods,” *IEEE International Conference on Multimedia Computing and Systems (ICMCS)*, pp. 90–95, 2012.
- [26] M. Sumithra and A. Devika, “A study on feature extraction techniques for text independent speaker identification,” *IEEE International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–5, 2012.
- [27] I. Trabelsi and D. Ben Ayed, “On the use of different feature extraction methods for linear and non linear kernels,” *6th IEEE International*

- Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pp. 797–802, 2012.
- [28] S. S. Nidhyananthan and R. S. S. Kumari, “Language and text-independent speaker identification system using GMM,” *WSEAS Transactions on Signal Processing*, vol. 9, no. 4, pp. 185–194, 2013.
  - [29] M. Moinuddin and A. N. Kanthi, “Speaker identification based on GFCC using GMM,” *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, vol. 1, no. 8, 2014.
  - [30] N. Almaadeed, A. Aggoun, and A. Amira, “Speaker identification using multimodal neural networks and wavelet analysis,” *IET Biometrics*, vol. 4, no. 1, pp. 18–28, 2015.
  - [31] Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, and M. Iwahashi, “Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–13, 2015.
  - [32] P. Kenny, “Joint factor analysis of speaker and session variability: Theory and algorithms,” *CRIM, Montreal, (Report) CRIM-06/08-13*, vol. 215, 2005.
  - [33] D. Matrouf, N. Scheffer, B. G. Fauve, and J.-F. Bonastre, “A straightforward and efficient implementation of the factor analysis model for speaker verification,” *Interspeech*, pp. 1242–1245, 2007.
  - [34] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of interspeaker variability in speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008.
  - [35] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, “An I-vector extractor suitable for speaker recognition with both microphone and telephone speech,” *Odyssey*, p. 6, 2010.
  - [36] P. Kenny, “A small footprint I-vector extractor,” *Odyssey*, pp. 1–6, 2012.
  - [37] P. Verma and P. K. Das, “I-Vectors in speech processing applications: a survey,” *International Journal of Speech Technology*, vol. 18, no. 4, pp. 529–546, 2015.

- [38] D. A. Reynolds, “Automatic speaker recognition using Gaussian mixture speaker models,” in *The Lincoln Laboratory Journal*. Citeseer, 1995.
- [39] —, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Communication*, vol. 17, no. 1, pp. 91–108, 1995.
- [40] D. Reynolds, “Large population speaker identification using clean and telephone speech,” *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46–48, 1995.
- [41] D. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.
- [42] V. R. Apsingekar and P. L. De Leon, “Support vector machine based speaker identification systems using GMM parameters,” *IEEE Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, pp. 1766–1769, 2009.
- [43] A. Revathi, R. Ganapathy, and Y. Venkataramani, “Text independent speaker recognition and speaker independent speech recognition using iterative clustering approach,” *International Journal of Computer science & Information Technology (IJCSIT)*, vol. 1, no. 2, pp. 30–42, 2009.
- [44] S. Bhardwaj, S. Srivastava, M. Hanmandlu, and J. Gupta, “GFM-based methods for speaker identification,” *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 1047–1058, 2013.
- [45] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer, and Y. Lei, “Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6773–6777, 2013.
- [46] T. Liu, K. Kang, and S. Guan, “I-vector based text-independent speaker identification,” *11th World Congress on Intelligent Control and Automation (WCICA)*, pp. 5420–5425, 2014.

- [47] L. Schmidt, M. Sharifi, and I. Lopez Moreno, “Large-scale speaker identification,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1650–1654, 2014.
- [48] R. Karadaghi, H. Hertlein, and A. Ariyaeinia, “Effectiveness in open-set speaker identification,” *International Carnahan Conference on Security Technology (ICCST)*, pp. 1–6, 2014.
- [49] P. Matjka, O. Glembek, O. Novotn, O. Plchot, F. Grzl, L. Burget, and J. H. Cernock, “Analysis of DNN approaches to speaker identification,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5100–5104, 2016.
- [50] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O’Leary, and B. A. Carlson, “The effects of telephone transmission degradations on speaker recognition performance,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 329–332, 1995.
- [51] D. A. Reynolds, “The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 113–116, 1996.
- [52] J. Ming, T. J. Hazen, J. R. Glass, and D. Reynolds, “Robust speaker recognition in noisy conditions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.
- [53] A. Khanteymoori, M. Homayounpour, and M. Menhaj, “Speaker identification in noisy environments using dynamic Bayesian networks,” *14th International CSI Computer Conference (CSICC)*, pp. 601–606, 2009.
- [54] N. Wang, P. Ching, N. Zheng, and T. Lee, “Robust speaker recognition using denoised vocal source and vocal tract features,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 196–205, 2011.
- [55] Q. Li and Y. Huang, “An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1791–1801, 2011.

- [56] X. Zhao and D. Wang, “Analyzing noise robustness of MFCC and GFCC features in speaker identification,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7204–7208, 2013.
- [57] H. Maged, A. AbouEl-Farag, and S. Mesbah, “Improving speaker identification system using discrete wavelet transform and AWGN,” *5th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 1171–1176, 2014.
- [58] X. Zhao, Y. Wang, and D. Wang, “Robust speaker identification in noisy and reverberant conditions,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.
- [59] M. A. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, “A robust speaker identification system using the responses from a model of the auditory periphery,” *PloS one*, vol. 11, no. 7, pp. 1–21, 2016.
- [60] Y. Hu, D. Wu, and A. Nucci, “Fuzzy-clustering-based decision tree approach for large population speaker identification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 762–774, 2013.
- [61] S. S. Nidhyananthan and R. S. S. Kumari, “A framework for multilingual text-independent speaker identification system,” *Journal of Computer Science*, vol. 10, no. 1, pp. 178–189, 2014.
- [62] S. Nandyal, S. S. Wali, and S. M. Hatture, “MFCC based text-dependent speaker identification using BPNN,” *International Journal of Signal Processing Systems*, vol. 3, no. 1, pp. 30–34, 2015.
- [63] S. S. Nidhyananthan, R. S. S. Kumari, and G. Jaffino, “Improving speaker identification performance by combining vocal tract features,” *International Journal of Applied Information Systems (IJ AIS)*, vol. 3, no. 1, pp. 27–33, 2012.
- [64] S. Nakagawa, L. Wang, and S. Ohtsuka, “Speaker identification and verification by combining MFCC and phase information,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1085–1095, 2012.

## REFERENCES

---

- [65] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, “TIMIT Acoustic-Phonetic Continuous Speech Corpus,” *Linguistic Data Consortium*, 1993. [Online]. Available: <https://catalog.ldc.upenn.edu/ldc93s1>
- [66] W. Fisher, G. Doddington, K. Goudie-Marshall, C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, “NTIMIT,” 1993. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC93S2>
- [67] E. B. George, K. L. Brown, M. Birnbaum, and M. Macon, “CTIMIT,” *Linguistic Data Consortium, Philadelphia, USA*, 1996. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC96S30>
- [68] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, “FFMTIMIT,” *Linguistic Data Consortium, Philadelphia, USA*, 1996. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC96S32>
- [69] D. Reynolds, “HTIMIT,” *Linguistic Data Consortium, Philadelphia, USA*, 1998. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC98S67>
- [70] A. Wrench, “MOCHA-TIMIT,” *Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, speech database*, 1999. [Online]. Available: <http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html>
- [71] C. Sanderson, “The VidTIMIT database,” IDIAP, Tech. Rep., 2002. [Online]. Available: <http://publications.idiap.ch/downloads/reports/2002/com02-06.pdf>
- [72] N. Morales, “STC-TIMIT,” *Linguistic Data Consortium, Philadelphia, USA*, 2008. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC2008S03>
- [73] P. Bauer and T. Fingscheidt, “WTIMIT,” *Linguistic Data Consortium, Philadelphia, USA*, 2010. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC2010S02>
- [74] N. Harte and E. Gillen, “TCD-TIMIT: An audio-visual corpus of continuous speech,” *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015. [Online]. Available: <https://sigmedia.tcd.ie/TCDTIMIT/>



## REFERENCES

---

- [75] A. Abdulaziz and V. Kepuska, “Noisy TIMIT Speech,” *Linguistic Data Consortium, Philadelphia, USA*, 2017. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC2017S04>
- [76] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, “Study of statistical robust closed set speaker identification with feature and score-based fusion,” *IEEE Statistical Signal Processing Workshop (SSP)*, pp. 1–5, 2016.
- [77] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, “DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM NIST speech NASA STI/Recon technical report N, vol. 93, pp. 1–94, 1993. [Online]. Available: [https://perso.limsi.fr/lamel/TIMIT\\_NISTIR4930.pdf](https://perso.limsi.fr/lamel/TIMIT_NISTIR4930.pdf)
- [78] C. Jankowski, “The NTIMIT speech database,” *printed documentation which accompanies the NTIMIT CD-ROM*, 1991. [Online]. Available: <https://catalog.ldc.upenn.edu/docs/LDC93S2/ntimit.readme.html>
- [79] “The Speakers In The Wild (SITW) speaker recognition challenge,” 2016. [Online]. Available: <http://www.speech.sri.com/projects/sitw/>
- [80] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, “The Speakers In The Wild (SITW) speaker recognition database.” *Interspeech*, pp. 818–822, 2016.
- [81] M. McLaren, A. Lawson, L. Ferrer, D. Castan, and M. Graciarena, “The speakers in the wild speaker recognition challenge plan,” 2015.
- [82] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, “The 2016 speakers in the wild speaker recognition evaluation.” *Interspeech*, pp. 823–827, 2016.
- [83] Y. Liu, Y. Tian, L. He, and J. Liu, “Investigating various diarization algorithms for Speaker In The Wild (SITW) speaker recognition challenge,” *Interspeech 2016*, pp. 853–857, 2016.
- [84] “2008 NIST Speaker Recognition Evaluation Training Set Part 2.” [Online]. Available: <https://catalog.ldc.upenn.edu/LDC2011S07>
- [85] “Free SFX Premium.” [Online]. Available: <http://www.freesfx.co.uk/>

## REFERENCES

---

- [86] “Find Sounds Search the Web for Sounds.” [Online]. Available: <http://www.findsounds.com/>
- [87] “IDIAP databset distribution portal.” [Online]. Available: <https://www.idiap.ch/dataset/mobio>
- [88] C. McCool and S. Marcel, “MOBIO Database for the ICPR 2010 face and speech competition,” Idiap, Tech. Rep., 2009.
- [89] J. Barker, M. Cooke, S. Cunningham, and X. Shao, “The GRID audiovisual sentences corpus,” 2007. [Online]. Available: <http://spandh.dcs.shef.ac.uk/gridcorpus/>
- [90] K. MacLean, “VoxForge,” *Ken MacLean*. [Online]. Available: <http://www.voxforge.org/home>. [Acedido em 2012].
- [91] J. Campbell and A. Higgins, “YOHO speaker verification,” *Linguistic Data Consortium, Philadelphia*, 1994. [Online]. Available: <https://catalog ldc.upenn.edu/ldc94s16>
- [92] J. P. Campbell and D. A. Reynolds, “Corpora for the evaluation of speaker recognition systems,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 829–832, 1999.
- [93] D. E. Sturim, P. A. Torres-Carrasquillo, and J. P. Campbell, “Corpora for the evaluation of robust speaker recognition systems,” *Interspeech 2016*, pp. 2776–2780, 2016.
- [94] L. Feng and L. K. Hansen, “A new database for speaker recognition,” Tech. Rep., 2005. [Online]. Available: [http://www2.imm.dtu.dk/pubdb/views/edoc\\_download.php/3662/pdf/imm3662.pdf](http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3662/pdf/imm3662.pdf)
- [95] H. Melin, “Databases for speaker recognition: Activities in COST250 working group 2,” *Final Report COST 250-Speaker Recognition in Telephony*, 1999. [Online]. Available: <http://www.speech.kth.se/prod/publications/files/1176.pdf>
- [96] “NIST online I-vector evaluations.” [Online]. Available: <https://ivectorchallenge.nist.gov/>

- [97] K. Nis, “Audio database toolbox,” *MATLAB, MANUAL*, vol. 21, 2009.  
[Online]. Available: <https://uk.mathworks.com/matlabcentral/fileexchange/23843-matlab-audio-database-toolbox>
- [98] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Springer Science & Business Media, 2007.
- [99] N. Dehak, “Discriminative and generative approaches for long-and short-term speaker characteristics modeling: application to speaker verification,” Ph.D. dissertation, École de technologie supérieure, 2009.
- [100] S. S. Nidhyananthan, R. Kumari, and G. Jaffino, “Robust speaker identification using vocal source information,” *IEEE International Conference on Devices, Circuits and Systems (ICDCS)*, pp. 182–186, 2012.
- [101] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, “Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification,” *4th IEEE International Conference on Biometrics and Forensics (IWBF)*, pp. 1–6, 2016.
- [102] S. S. Yadav and D. Bhalke, “Speaker identification system using wavelet transform and VQ modeling technique,” *International Journal of Computer Applications*, vol. 112, no. 9, pp. 19–23, 2015.
- [103] V. R. Apsingekar and P. L. De Leon, “Speaker model clustering for efficient speaker identification in large population applications,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 848–853, 2009.
- [104] C. Kim and R. Stern, “Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring,” *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 4574–4577, 2010.
- [105] K. Kumar, C. Kim, and R. M. Stern, “Delta-spectral cepstral coefficients for robust speech recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4784–4787, 2011.

## REFERENCES

---

- [106] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.
- [107] E. Gopi, *Digital Speech Processing using Matlab*. Springer, 2014.
- [108] T. Herbig, F. Gerl, and W. Minker, *Self-Learning Speaker Identification: A System for Enhanced Speech Recognition*. Springer Science & Business Media, 2011.
- [109] F. E. A. El-Samie, "Information security for automatic speaker identification," *Information Security for Automatic Speaker Identification*, pp. 1–122, 2011.
- [110] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, Jan 2006.
- [111] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," *Speech Communication*, vol. 49, no. 6, pp. 501–513, 2007.
- [112] S. Singh, M. H. Assaf, S. R. Das, S. N. Biswas, E. M. Petriu, and V. Groza, "Short duration voice data speaker recognition system using novel fuzzy vector quantization algorithm," *IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pp. 1–6, May 2016.
- [113] Y. Suh and H. Kim, "Discriminative likelihood score weighting based on acoustic-phonetic classification for speaker identification," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 126, 2014.
- [114] M. J. Alam, V. Gupta, P. Kenny, and P. Dumouchel, "Speech recognition in reverberant and noisy environments employing multiple feature extractors and I-vector speaker adaptation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 50, 2015.
- [115] C. S. Kumar and P. M. Rao, "Design of an automatic speaker recognition system using MFCC, vector quantization and LBG algorithm," *International*

- Journal on Computer Science and Engineering*, vol. 3, no. 8, pp. 2942–2954, 2011.
- [116] A. A. Ross, K. Nandakumar, and A. Jain, *Handbook of Multibiometrics*. Springer Science & Business Media, 2006.
  - [117] A. Ross and A. Jain, “Information fusion in biometrics,” *Pattern recognition letters*, vol. 24, no. 13, pp. 2115–2125, 2003.
  - [118] “SITW Database.” [Online]. Available: <http://www.speech.sri.com/projects/sitw/>
  - [119] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
  - [120] H. Misra, S. Iqbal, and B. Yegnanarayana, “Speaker-specific mapping for text-independent speaker recognition,” *Speech Communication*, vol. 39, no. 3, pp. 301–310, 2003.
  - [121] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.
  - [122] T. Stafylakis, P. Kenny, M. J. Alam, and M. Kockmann, “Speaker and channel factors in text-dependent speaker recognition,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 1, pp. 65–78, 2016.
  - [123] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, “Improving short utterance I-vector speaker verification using utterance variance modelling and compensation techniques,” *Speech Communication*, vol. 59, pp. 69–82, 2014.
  - [124] H. Zeinali, A. Mirian, H. Sameti, and B. BabaAli, “Non-speaker information reduction from cosine similarity scoring in I-vector based speaker verification,” *Computers & Electrical Engineering*, vol. 48, pp. 226–238, 2015.

## REFERENCES

---

- [125] A. Larcher, K. A. Lee, B. Ma, and H. Li, “Text-dependent speaker verification: Classifiers, databases and RSR2015,” *Speech Communication*, vol. 60, pp. 56–77, 2014.
- [126] J. M. K. Kua, J. Epps, and E. Ambikairajah, “I-Vector with sparse representation classification for speaker verification,” *Speech Communication*, vol. 55, no. 5, pp. 707–720, 2013.
- [127] S. Cumani, N. Brummer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, “Pairwise discriminative speaker verification in the I-vector space,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1217–1227, June 2013.
- [128] A. Fazel and S. Chakrabartty, “An overview of statistical pattern recognition techniques for speaker verification,” *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 62–81, 2011.
- [129] O. Ghahabi and J. Hernando, “Deep belief networks for I-vector based speaker recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1700–1704, 2014.
- [130] —, “I-vector modeling with deep belief networks for multi-session speaker recognition,” *network*, vol. 20, p. 13, 2014.
- [131] W. Rao and M.-W. Mak, “Boosting the performance of I-vector based speaker verification via utterance partitioning,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1012–1022, 2013.
- [132] O. Ghahabi and J. Hernando, “Deep learning for single and multi-session I-vector speaker recognition,” *arXiv preprint arXiv:1512.02560*, 2015.
- [133] L. Li, D. Wang, C. Zhang, and T. F. Zheng, “Improving short utterance speaker recognition by modeling speech unit classes,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 6, pp. 1129–1139, 2016.
- [134] S. Cumani, O. Plchot, and P. Laface, “On the use of I-vector posterior distributions in probabilistic linear discriminant analysis,” *IEEE/ACM*

- Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 4, pp. 846–857, 2014.
- [135] Z. Boulkenafet, M. Bengherabi, F. Harizi, O. Nouali, and C. Mohamed, “Forensic evidence reporting using GMM-UBM, JFA and I-vector methods: Application to Algerian Arabic dialect,” *8th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 404–409, Sept 2013.
- [136] A. Khodabakhsh, A. Mohammadi, and C. Demiroglu, “Spoofing voice verification systems with statistical speech synthesis using limited adaptation data,” *Computer Speech & Language*, vol. 42, pp. 20–37, 2017.
- [137] T. Hasan and J. H. Hansen, “Maximum likelihood acoustic factor analysis models for robust speaker verification in noise,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 2, pp. 381–391, 2014.
- [138] L. Chen and Y. Yang, “Emotional speaker recognition based on I-vector through atom aligned sparse representation,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7760–7764, 2013.
- [139] L. Macková, A. Čížmár, and J. Juhár, “Emotion recognition in I-vector space,” *26th IEEE International Conference Radioelektronika (RADIOELEKTRONIKA)*, pp. 372–375, 2016.
- [140] M. H. Bahari, M. McLaren, H. V. Hamme, and D. V. Leeuwen, “Age estimation from telephone speech using I-vectors,” *Proceedings Interspeech*, pp. 506–509, 2012.
- [141] W. B. Kheder, D. Matrouf, P.-M. Bousquet, J.-F. Bonastre, and M. Ajili, “Robust speaker recognition using MAP estimation of additive noise in I-vectors space,” *International Conference on Statistical Language and Speech Processing*, pp. 97–107, 2014.
- [142] K. P. Markov and S. Nakagawa, “Text-independent speaker recognition using non-linear frame likelihood transformation,” *Speech Communication*, vol. 24, no. 3, pp. 193–209, 1998.

- [143] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan, “Second-order statistical measures for text-independent speaker identification,” *Speech Communication*, vol. 17, no. 1-2, pp. 177–192, 1995.
- [144] X. Lu and J. Dang, “An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification,” *Speech Communication*, vol. 50, no. 4, pp. 312–322, 2008.
- [145] J.-Y. Lai, S.-L. Wang, A. W.-C. Liew, and X.-J. Shi, “Visual speaker identification and authentication by joint spatiotemporal sparse coding and hierarchical pooling,” *Information Sciences*, vol. 373, pp. 219–232, 2016.
- [146] K. Kumar, Q. Wu, Y. Wang, and M. Savvides, “Noise robust speaker identification using Bhattacharyya distance in adapted Gaussian models space,” *16th IEEE European Signal Processing Conference*, pp. 1–4, 2008.
- [147] Y. Goto, T. Akatsu, M. Katoh, T. Kosaka, and M. Kohda, “An investigation on speaker vector-based speaker identification under noisy conditions,” *IEEE International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 1430–1435, 2008.
- [148] M. Li and S. Narayanan, “Simplified supervised I-vector modeling with application to robust and efficient language identification and speaker verification,” *Computer Speech & Language*, vol. 28, no. 4, pp. 940–958, 2014.
- [149] M. Kim, E. Kim, C. Seo, and S. Jeon, “Speaker verification and identification using principal component analysis based on global eigenvector matrix,” *International Conference on Hybrid Artificial Intelligence Systems*, pp. 278–285, 2010.
- [150] P. Li, Y. Li, D. Luo, and H. Luo, “Speaker identification using FrFT-based spectrogram and RBF neural network,” *34th IEEE Chinese Control Conference (CCC)*, pp. 3674–3679, 2015.
- [151] K. Daqrouq, A.-R. Al-Qawasmi, W. Al-Sawalmeh, and T. A. Hilal, “Wavelet transform based multistage speaker feature tracking identification system using linear prediction coefficient,” *IEEE International Conference on Advances in Computational Tools for Engineering Applications ACTEA’09.*, pp. 173–179, 2009.



- [152] H. Zheng, M. Wang, and Z. Li, “Audio-visual speaker identification with multi-view distance metric learning,” *17th IEEE International Conference on Image Processing (ICIP)*, pp. 4561–4564, 2010.
- [153] D. N. Woo and R. S. Aygün, “Unsupervised speaker identification for TV news,” *IEEE MultiMedia*, vol. 23, no. 4, pp. 50–58, 2016.
- [154] N. M. AboElenein, K. M. Amin, M. Ibrahim, and M. M. Hadhoud, “Improved text-independent speaker identification system for real time applications,” *Fourth IEEE International Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC)*, pp. 58–62, 2016.
- [155] G. Huang, G.-B. Huang, S. Song, and K. You, “Trends in extreme learning machines: A review,” *Neural Networks*, vol. 61, pp. 32–48, 2015.
- [156] E. Cambria, G.-B. Huang, L. L. C. Kasun, H. Zhou, C. M. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li *et al.*, “Extreme Learning Machines [trends & controversies],” *IEEE Intelligent Systems*, vol. 28, no. 6, pp. 30–59, 2013.
- [157] Y. Lan, Z. Hu, Y. C. Soh, and G.-B. Huang, “An Extreme Learning Machine approach for speaker recognition,” *Neural Computing and Applications*, vol. 22, no. 3-4, pp. 417–425, 2013.
- [158] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: theory and applications,” *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.
- [159] L. L. C. Kasun, Y. Yang, G.-B. Huang, and Z. Zhang, “Dimension reduction with extreme learning machine,” *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3906–3918, 2016.
- [160] Z. Huang, Y. Yu, J. Gu, and H. Liu, “An efficient method for traffic sign recognition based on extreme learning machine,” *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 920–933, 2017.
- [161] G.-B. Huang, “An insight into extreme learning machines: random neurons, random features and kernels,” *Cognitive Computation*, vol. 6, no. 3, pp. 376–390, 2014.

## REFERENCES

---

- [162] —, “What are extreme learning machines? filling the gap between Frank Rosenblatt’s dream and John von Neumann’s puzzle,” *Cognitive Computation*, vol. 7, no. 3, pp. 263–278, 2015.
- [163] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, “Speaker identification evaluation based on the speech biometric and I-vector model using the TIMIT and NTIMIT databases,” *5th IEEE International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6, 2017.